Social Engagement at Scale

Kowalczyk, Damian Konrad

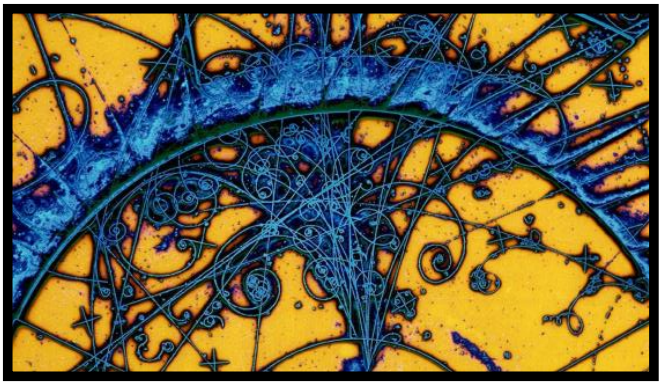
Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

PHD THESIS

SOCIAL ENGAGEMENT AT SCALE

DAMIAN KONRAD KOWALCZYK



April 2020, Denmark

Technical University of Denmark
Department of Applied Mathematics and Computer Science

Microsoft Corporation
Business Applications Group, Cloud + AI Division

Microsoft

DTU

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Microsoft Corporation
Business Applications Group, Cloud + AI Division
MDCC, Kanalvej 7
2800 Kongens Lyngby, Denmark
Phone +45 4567 8000
mdcc@microsoft.com
www.microsoft.com

front page illustration:
Particle tracks in the Big European Bubble Chamber (BEBC)
European Organization For Nuclear Research (CERN)
Geneva 23, Switzerland

To my mother, Maria.

Preface

The first seed for this project was planted in 2007, at the European Organization for Nuclear Research in Geneva. There, supervised by Samim Erhan and Alexander Oh, I have witnessed physicists accelerating and colliding particles in pursuit of fundamental knowledge. In the summer, we have been investigating why the Run Control System [PGK+07] for the Compact Muon Solenoid (CMS) Experiment [Pim90] would take a whole minute to initialize data acquisition. Many lessons resonate with me over a decade later:

- Particle collisions offer far more insight than observing individual particles at rest. High energy physics at CERN has so far seen investment exceeding €10 billion.
- There is knowledge that can only be acquired at scale. The thirteen thousand tons CMS detector records petabytes of data every second, but is only one of five experiments for the Large Hadron Collider.
- At scale is where science becomes painfully dependent on big data engineering. The principles of distributed systems engineering are still not part of the nuclear physics curriculum.
- The best solutions require an inter-disciplinary approach above old boundaries. There have been 80 nationalities present on the campus in 2017. No one questioned gender, race or neurodiversity, as if these paved the best way forward.

That summer, I have also received my first friend invite on Facebook, coming out of the exclusive academic trial. It took another decade until the US Presidential Elections in 2016, for both opportunities and threats of online social networks to finally receive the necessary spotlight. Today social networks aggregate and direct the attention of billions, directly exposing our wellbeing and the democratic process itself. The value they can deliver appears underestimated as well, and both physicists [Sco11, KMT19] and social scientists [CKK14, TBDH19] take note. Twitter alone can deliver a high-luminosity beam of 700 million events per day, known as the Firehose. Many of these events can be described as collisions. The mechanisms behind these events are not new, the opportunity to understand them is.

> "Maps of social networks are based on the degree of connectivity between people. It's the same idea here." (Jesse Thaler, 2019)

Raison d'être of this thesis, is pure stubborn optimism, primed by CERN. Because if physicists are right, then social scientists should also ask big questions, for the social colliders are here to deliver the answers. If nuclear physics is positioned to explain the universe, then computational social science is positioned to advance the human condition. The goal of this thesis is to motivate such optimism with numbers, delivered above two old boundaries: the one dividing industry and academia, and the one dividing engineering and science, as long as I am standing on the shoulders of giants.

On 4 July 2012, the Run Control for the CMS detector facilitated the discovery of the Higgs Boson. On 8 October 2013, Professor Higgs and Professor Englert have been awarded the Nobel Prize in Physics.

Kongens Lyngby, 22-April-2020

Damian Konrad Kowalczyk

Summary

This project examines the dynamics of human attention, in the times where so much of it is aggregated and commoditized by the online social networks. The digital townhalls of Twitter, Facebook and Instagram track our collective attention via an increasingly diverse set of engagement metrics. The first study, proves it is possible to advance the state-of-the-art in virality prediction without compromising on explainability, robustness or privacy compliance. The approach combines content signal available at the time of posting, with high accuracy ground-truth and sentiment analysis in 18 languages to achieve state-of-the-art results on multiple benchmark datasets. The second study, questions virality as the best predictor of social influence. We examine a diverse set of content engagement metrics from Twitter. Correlations discovered lead us to propose a new, more holistic, one-dimensional engagement signal. We then show it is more predictable than any individual influence metric previously investigated. We propose the ability to engage the audience as a new, more holistic target for social influence maximization and share the compound engagement workflow to ensure reproducibility. In the third study, we examine the transferability of the proposed framework beyond Twitter. We address the problem of multi-modal popularity prediction on Instagram. We use deep neural networks to advance user-generated content representation. Through the ablation of transfer learning, we offer a detailed explanation of popularity dynamics. The models of virality, engagement and popularity are the first to achieve strong ranking performance in a robust and explainable way. The compound engagement model, in particular, is the first to explain half of the variance with features available early, and to offer strong ranking performance simultaneously. I deliver new models of understanding, via scientific avenues and Microsoft cloud services.

The era of big data offers significant advancements in data collection, storage and analysis methods, creating new opportunities for researchers to achieve high relevance and impact. Extracting knowledge from social big data, however, remains extremely difficult. Much of the recent work is still plagued by anecdotal evidence from short timeframe samples or black-box approaches, while the relevant technology, data and knowledge appear siloed in separation. One ambition of this Industrial PhD project is to rise above the divide between engineering and science and prove the potential of a holistic approach. The proposed data collection and analysis framework positioned this project among the largest studies on social media to date. The proposed model operationalization framework enabled Microsoft customers to respond to pre-viral content, including support request, before anyone else.

Dansk resume

Dette projekt undersøger dynamikken i den menneskelige opmærksomhed i en tid, hvor så meget af den er samlet og kommercialiseret af de digitale hubs på Twitter, Facebook og Instagram. Online sociale netværk sporer vores kollektive opmærksomhed gennem stadig flere forskelligartede typer af engagementsmålinger. Den første undersøgelse viser, at det er muligt at forbedre state-of-the-art forudsigelser om hvordan indhold går viralt uden at gå på kompromis med forklarbarhed, robusthed eller overholdelse af privatlivets fred. Metoden kombinerer features, der er tidligt tilgængelige med en nøjagtig ground-truth og stemningsanalyse på 18 sprog for at opnå avancerede resultater på flere benchmarkdatasæts. I den anden undersøgelse sætter vi spørgsmålstegn ved hvorvidt viralitet er den bedste indikator for social indflydelse. Vi undersøger et forskelligartet sæt af indholdsengagementmålinger fra Twitter. De fundne korrelationer fører til, at vi foreslår et nyt og mere holistisk og en-dimensionelt engagementssignal. Vi viser derefter, at det er mere forudsigeligt end nogen tidligere undersøgt individuel indflydelsesparameter. Vi foreslår muligheden for at engagere publikum som et nyt og mere holistisk mål for maksimal social indflydelse samt deler det sammensatte engagements-workflow for at sikre reproducerbarhed. I den tredje undersøgelse beviser vi, at de foreslåede rammer kan overføres ud over Twitter. Vi løser problemet med multimodal popularitetsforudsigelse på Instagram. Vi bruger dybe neurale netværk til at videreudvikle brugergenereret indholdsrepræsentation. Gennem ablering af transferlearning fremlægger vi en detaljeret forklaring på popularitetsdynamik. Modellerne for viralitet, engagement og popularitet er de første, som opnår en stærk rankingperformance på en robust og forklarlig måde. Særligt er den sammensatte engagementsmodel den første, der forklarer halvdelen af variansen med features, der er tilgængelige tidligt, og samtidig tilbyder en stærk rankingpræstation. Jeg leverer nye forståelsesmodeller ved brug af videnskabelige metoder og Microsoft cloud-tjenester.

Big Data-æraen giver betydelige fremskridt inden for dataindsamling, opbevaring og analysemetoder, hvilket skaber nye muligheder for forskere til at opnå høj relevans og virkning. Det er dog fortsat ekstremt vanskeligt at udtrække viden fra sociale big data. Meget af det nylige arbejde er stadig plaget af anekdotisk dokumentation fra korte tidsrammeeksempler eller black-box-tilgange, mens den relevante teknologi, data og viden fremstår adskilt. En af ambitionerne med dette erhvervs-ph.d.-projekt er at hæve sig over adskillelsen mellem ingeniørvidenskab og videnskab samt bevise potentialet i en helhedsorienteret tilgang. Den foreslåede dataindsamlings- og analyse-ramme har placeret dette projekt blandt de mest omfangsrige studier af sociale medier. Den foreslåede operationelle model har gjort det muligt for Microsoft kunder at reagere på præ-viralt indhold, inklusiv supportanmodning, før nogle andre.

Contributions

Papers and manuscripts included in this thesis:

- Paper A: Damian Konrad Kowalczyk and Jan Larsen. Scalable Privacy-Compliant Virality Prediction on Twitter. In *Proceedings of AffCon 2019 @ AAAI*, volume 2328, pages 12–27, 2019. (Best Paper Award)
- Paper B: Damian Konrad Kowalczyk and Lars Kai Hansen. The complexity of social media response: Statistical evidence for one-dimensional engagement signal in twitter. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, 2020
- Paper C: Christoffer Riis, Damian Konrad Kowalczyk, and Lars Kai Hansen. On the limits to multi-modal popularity prediction on instagram – a new robust, efficient and explainable baseline. 2020. (Submitted)

Acknowledgments

This project is supported by the Business Applications Group within Microsoft and the Danish Innovation Fund, Case No. 5189-00089B. I want to thank my principal supervisors Lars Kai Hansen and Jörg Derungs, who trusted me with their time and support long before any quantifiable justification. When I stand tall it is because of their shoulders. I would like to thank the Business Applications Group and specifically Pushpraj Shukla, Ralf Gautschi, Uffe Kjall, Sandeep Aparajit and Walter Sun for their substantial and indispensable support offered throughout the project.

I owe much gratitude to Charlotte Mark for all the help during market research, grant application and IP negotiations. I remain thankful to my supervisors at CERN, Samim Erhan and Alexander Oh, for giving me the best possible introduction to interdisciplinary science at scale, a key inspiration for this project. I will never forget who believed in me early. Special thanks go to my friends and mentors for the abundance of experience and common sense shared with me over the years, including (alphabetically): Peter Christensen, Jörg Derungs, Krystyna and Tomasz Janiczek, Nicholas Fish, Barbara Lechner, Signe Lund, Paweł Kruk, Piotr Madejski, Michael Neuburger, Michał Sroka, Katarzyna and Łukasz Olbromski, Anna and Gaurav Roy, Mayank Shrivastava, Steen Tommerup and Tomasz Truszkiewicz.

Finally, I want to recognize the estimated 5-15% of our society, born on the neurodiverse spectrum, naturally predisposed to innovate across disciplines. Many of them are still marginalized, discouraged from reaching their full potential, distracted by the modern media and consequently unaware of their unique strengths in the times of AI [WS08, AP17].

Contents

Chapter 1

Introduction

"Whoever treats of interest inevitably treats of attention."
(The Principles of Psychology, William James)

1.1 The Attention Economy

It has been argued for centuries, that "in a free and open encounter" of ideas, truth will prevail [Dze00, QFS+17]. John Milton's argument was used by Oliver Holmes, to propose the foundation of the free marketplace of ideas: the assumption that the best test of truth is the power of an idea to get itself accepted in the competition of the market. This assumption remains central in free speech thought until today [HS15b]. The concept of a free marketplace of ideas has since then been used to support free speech policy, modern economics, adoption of online social platforms, and applied to the study of scientific research itself [MGTW12, QFS+17]. [QFS+17] highlights two necessary elements of Holmes's theory for the success of the marketplace: the diversity of ideas to which people are exposed and the discriminative power of the marketplace, which they define as the ability to allow better ideas to become more popular. Online social networks (OSNs) are of increasing significance in the context of the free marketplace of ideas, for the abundance and diversity of information.

The digital townhalls of Twitter and Facebook show a significant impact on the marketplace by broadening participation and facilitating the ever-increasing flow and exchange of "ideas", "posts", or "memes", among many other transmissible pieces of information [Daw76, QFS+17]. However, the discriminative power of the networks is questioned [QFS+17, VBZ+16, FVD+16]. In fact the recent years of globalizing the marketplace of ideas appear to face a thorough of disillusionment, as the OSNs far exceed the limits of our collective attention, becoming the preferred medium of communication for individuals and organisations around the world [KL19, KH20]. Through networks such as Twitter, Facebook and Instagram, users today are exposed to an unprecedented volume of information, competing for their engagement. Human's cognitive constraints, however, limit our collective attention, and the number of social interactions we can sustain, or the number of ideas we can consider [GPV11, PBC12, QFS+17, LSMHL19]. These constraints give rise to the "attention economy" [HKS71, Gol97, Fal07, CFM15, Ter12, Fra19, vK19, QFS+17].

> **Attention Economy** "a system that revolves primarily around paying, receiving, and seeking what is most intrinsically limited and not replaceable by anything else, namely the attention of other human beings" (Michael Goldhaber, 2006)

In theories of the attention economy, attention is, first of all, a scarce resource. The scarcity of attention, allows the Internet to become an economical medium again, that is, a medium to which all the axioms of market economics can once again be applied [Ter12]. As explained by [vK19], there are four core elements to note in Franck's new theory: the fundamental human desire for attention; the parallel between attention and money, the self-reproducing character of attention capital (due to preferential attachment) earning interest just as money does; and the pervasive desire of everyone becoming a celebrity and a 'brand' within the new 'mental capitalism' [vK19, Fra99, Fra19].

The centrality of the notion of attention, stands in a radical contrast, to the centrality of information in earlier theorisations of the economy of the Internet and digital media [Gol06, Bar93, Kel99]. The latter still guides the industrial practice of the big data aggregators until today [MSC13, Jam11, GHTA18, FB13].

> "In an earlier phase, new media economists stress the abundance of information in the digital economy, to assert a new kind of economic Darwinism, based on the capacities of a proliferating, connected life to create the new. [...] The bios of the attention economy, however entails a continuity with the Darwinian dynamics of competition, within the harsh constraints of natural scarcity which framed the notion of the survival of the fittest." (Terranova, 2012)

The definition and measure of fitness, meanwhile, have evolved dramatically. Today, the digital media form a system of channels supplying information in order to extract attention, and our affluent society ranks income in attention above financial success, as argued by Franck 20 years later [Fra19]. Consequently, social influence is now built on attention capital and measured as the ability of an author to spread information in their networks [PAP+13, EMR+19].

Over recent years, the attempts to capitalise attention have gone far [Wu17]. Platforms like Facebook and Twitter, offer an unparalleled opportunity for influence analysis and maximisation, impacting public opinion, culture, policy, and commerce [DB01]. The abstract quality of attention, aggregated by platforms like Facebook, Twitter and Instagram, combined with automated forms of measurement (via 'likes', 'comments', 'views', 'followers' or 'retweets' of user-generated content) have opened our collective attention to marketization and financialization [Ter12]. The market capitalisation of the resulting industry, based on aggregating, predicting and directing users attention, is now measured in trillions of dollars.

The unprecedented amount of attention aggregated by OSNs comes under intense criticism in the recent years [Bue16, Wu17, Bey19, BJ19], as billions are now exposed to low-quality content and questionable influence [QFS+17, KH20]. Our behavioural mechanisms to cope with information exceeding our limited attention, make these information markets less meritocratic and diverse, increasing the spread of misinformation [NR10, VBZ+16] and making us vulnerable to manipulation [FVD+16, RCM+11, SRN+18a]. Automatic recommendation algorithms may cluster people into a few homogeneous factions [Axe97], often called "echo chambers" [Sun14] or "filter bubbles" [Par11]. The abundance of anecdotal evidence of hoaxes, conspiracy theories, and fake news in online social media renders massive digital misinformation among the top global risks for our society today [WJC+15].

To introduce this project through the lens of limitless possibilities of the digital age and abundant prior attempts to materialise it, would be shortsighted if not naive. The scarcity of attention and asymmetry of access define this project from day one. In the world where the struggle for attention replaces the struggle for material goods, understanding the dynamics of attention in the online markets of ideas becomes more important than ever.

1.2 The Dynamics of Attention

In the times where social influence is built on attention capital [EMR+19, Fra19], the amount of attention received by the user-generated content (UGC) becomes predictive of her influence. Online Social Networks measure this attention via an increasingly diverse set of UGC engagement metrics. Some of them remain in the focus of high activity in the field until today. This section offers a summary of prior attempts to understand and predict them.

> "The role of the social and professional networks in the spread and acceptance of innovations, knowledge, business practices, products, behaviour, rumours, and memes, is a much-studied problem in social sciences, marketing and economics. Online environments like Twitter offer an unprecedented opportunity to track such phenomena. Consequently, a staggering number of studies focus on social spreading, asking, for example, why can some messages reach millions of individuals, while others struggle to get noticed."
>
> (Albert-László Barabási, 2016)

1.2.1 User Generated Content Virality

Extant work on influence analysis focuses on homogeneous information networks and attributes the greatest influence to authors triggering the largest diffusion cascades [Fra19, KH20]. On Twitter, the size of these cascades is measured by the retweet count. This metric is often assumed as a measure of a tweet's popularity, or more precisely virality, and since [PAP+13] considered as the best predictor of UGC influence.
[TLP14] show that carefully crafted wording of the message could help propagate the tweets better, however, the 140-character constraint imposed by Twitter, makes it difficult to identify and extract predictive features [COM13], challenging text-based popularity prediction. However, there is much more to Twitter's UGC than the caption. [IKT12, WBF18] demonstrate the much higher performance of network-oriented features (e.g., number of followers) in predicting image popularity on Twitter. [PAP+13, KLPM10, CHBG10] demonstrate relationships between the user's follower count and their influence on information spreading in the network. Ranking users by the count of followers is found to perform similarly to PageRank [KLPM10]. [PAP+13, COM13] note retweet count follows a power law, and model the probability to be retweeted after log transformation. [PDB13] predicts a range of the logarithm of expected retweets, via Random Forest classification. He shows the predictive value of

user features (e.g., count of followers), network features, and the popularity of hashtags included. [BT16] provide a comparison of learning methods and features for virality prediction. They show Random Forests to achieve the best performance in the binary classification of virality and highlight the value of author features again: number of followers, the number of times a user is listed by others, and the average number of tweets posted per day. [NPPZ18] employs recursive partitioning trees to achieve 0.682 classification accuracy on a large topical dataset, however using features unavailable early (favourites count) or not available anymore (local publication time, discontinued by Twitter in 2019), challenging both scalability and reproducibility. [HAN+11] investigated the features of tweets contributing to retweetability and are the first to explore the effect of negative sentiment on the diffusion of news on Twitter.
Substantial gains in predictive performance are seen after including network features calculated from the content graph formed by retweets, or the relationship graph formed by friendship. In order to deliver such signal, the document level subgraphs are often built via real-time monitoring of the diffusion process [KL19]. [ZHvGS10] predicted the popularity of a tweet via time-series path of its retweets using a Bayesian probabilistic model. [WBF18] uses a preconditioned recurrent neural network to model the temporal diffusion and shows a ranking performance of 0.366 on benchmark datasets. [ASH13] used temporal evolution patterns to predict the popularity of online UGC. [ZEH+15] model the retweeting cascades as a self-exciting point process. [FDS16] shows it's possible to boost predictive accuracy after determining the topics of interest of the author based on his past tweets. [PZP+11] studied retweet network propagation trends using conditional random fields, demonstrating gains in accuracy when considering social relationships and retweet history together. However, access to subgraphs on the author or even document level is strictly rate-limited by social networks [KL19]. Therefore leveraging tweet's early performance, author's relationships, preferences or retweet history is prohibitive for a scalable, near real-time prediction on a single tweet [KL19].

1.2.2 Multimodal Popularity Prediction

The complexity and urgency of UGC popularity prediction continue inspiring high activity in the field. Researchers around the world move to deep multimodal UGC representation in attempts to advance state-of-the-art. Their attention and analytic workload shifts to building the best UGC representation ahead of predictive analysis. This multimodal representation is now based on all UGC modalities at hand, such as metadata, user information, text (UGC caption) and visual attachments. The various attempts at extracting relevant features from image attachments, in particular, benefit greatly from the high activity in the field of computer vision. Deep Neural Network (DNN) architectures continue

advancing variety of relevant tasks: image classification [TL19a, SVZ13, RW17, CMS12], object detection [RF18, CFFV16, ESTA14, STE13] or image segmentation [Bto. . . 17, BLSA17, CPK$^+$14, HDWF$^+$17]. Recurrent and otherwise Deep Neural Networks consequently advance feature extraction from high-dimensional unstructured image attachments (mandatory on Instagram). There have been at least five contributions in UGC popularity prediction, to benefit from DNNs in last October alone [CLZ$^+$19, DMW19, DWW19, HHWY19, HLK$^+$19]. These advancements regularly come at the cost of scalability or generalizability. Early works show promising results using the visual modality. [CMS15] shows performance boost using only visual features extracted from a single pre-trained DNN, which already motivates, including visual modality as a predictor of UGC popularity. The experiments of [KDH14] show how simple image statistics do not improve the performance, but that both low-level features (e.g., gist, texture, colour patches, gradient or DNN embeddings) combined with semantic features, e.g., describing the detected objects, lead to significant performance gains. [KDH14] concludes that detecting scenes, objects and faces are relevant for predicting image popularity. [MMJ14] consider colour features, scene classification, face detection together. [CAD$^+$14] use temporal and structural features to predict the size of diffusion cascade for an image on Facebook. [MMJ14] used textual, visual and social cues to predict the image popularity on Flickr. [WBF18] proposed a joint-embedding DNN which combines the same cues to rival SOTA on Twitter. Both [MPW18, MRR$^+$16] show that dividing posts into categories and brand-related concepts offers a boost to performance, but at the expense of generality. [GO19] shows how elaborate feature extraction from several sources, combined with embeddings of past posts yields impressive results, however again, at the expense of scalability and generality. Multiple recent advancements of the state-of-the-art in popularity ranking performance [CKXM19, WBF18] are in fact due to multimodal representation. However, the predictive gains extracted from the visual modality alone were relatively small so far [DMW19, HHWY19, HLK$^+$19, DWW19]. Simple cost-to-benefit calculation of employing DNNs would challenge their choice in production scenarios. [HHWY19] uses word embeddings extracted from the textual modality with word2vec. They use LightGBM [KMF$^+$17] for predictive modelling and perform ablation study, which shows that the textual modality improves model performance. [HHWY19], show however that visual features extracted with ResNet-152 [HZRS16a] did not help improving model's performance. [WBF18] employs all four modalities (metadata, user information, textual and visual information) by constructing a late-fused joint embedding, which is then used in modelling retweet count in Poisson regression. Via ablation study, they show that the metadata and user information offers the most important features, but the best performance requires using all four modalities. [HHWY19, WBF18] show that naively extracting and concatenating features from different modalities, does not immediately improve predictive performance, however after extensive feature engineering the performance boost was significant.

1.3 Knowledge Extraction and Delivery at Scale

Online traces of human activity offer novel opportunities to study the dynamics of attention, and in particular how emergent patterns of collective attention determine what new information is generated and consumed [CFM15, MLCM13, MCA+13, CS08, CKK14]. The gathering, fusion, processing and analysing of the big social media data from unstructured (or semi-structured) sources make knowledge extraction an extremely difficult task which has not been completely solved [KAEM13, BOJC16, KSN20]. Classic data management methods, algorithms, frameworks and tools have become inadequate for processing the vast amount of data, as they do not properly scale while the data size increases [BOJC16]. It is perhaps ironic, to watch their limits exceeded in the same fashion as those of our collective attention. As noted by [TBDH19], tools which can help the analyst to reason using more data or less biased data are not widely used, are often more complex than the average analyst wants to use, or require more (e.g., time) than the analyst wants to spend to generate results. Together with [BOJC16] they argue the need for more scalable technologies, but also for a better understanding of the biases in the data and ways to overcome them. This section aims to introduce the most relevant advancements and emerging standards in knowledge extraction and delivery at scale.

1.3.1 Social Big Data Analysis

Social big data (SBD) has become essential for various distributed services, applications, and systems [PYM18], enabling or accelerating event detection [DMCF15], sentiment analysis [Fel13, FZ15, HBA15], popularity prediction [WS15, CMS15, DMW19, WBF18], viral marketing [Dom05, RCC20], influencer identification [AW11a, AW11b, WZS+16, BKA09], personalized recommendation [GJ18], online advertising [BP08], opinion leader detection [BNQ19] and many other [CKK14, BOJC16]. Computational and storage requirements of such applications lead to the cloud-scale reinvention of data storage and processing technologies. New tools are emerging to replace the conventional non-effective ones, and a hybrid of techniques is now necessary for knowledge extraction at scale [KAEM13, GH15, KL19]. The driving forces behind the development of new tools are perhaps best understood through the core concepts of big data, initially summarised with the 3V model by Laney [Lan01] and refined by Gartner [BL12]:

> **3Vs of Big Data** "the high volume, high velocity, and/or high variety information assets that require new forms of processing to enable

> enhanced decision making, insight discovery and process optimisation" (Beyer, 2012)

The arrival of social big data makes the knowledge discovery process ever more tangled. The volume, velocity, and variety of mostly unstructured information from a single social network are evolving rapidly, while the social nature of nodes in these networks makes data subjective to many privacy concerns and laws [LBHW20, KL19]. The additional challenges motivate an extension of the 3V model to 5V [HYA+15, BOJC16], and receive a comprehensive analysis by [BOJC16] covering achievements until 2015. Below offers a 5V summary of the recent challenges and advancements most relevant for this project.

- **Volume:** 700 million tweets and 95 million images have been posted on Twitter and Instagram on the day of writing this section. The volume generated by these platforms is now measured in petabytes per day, demanding cloud-scale storage and compute technologies to begin knowledge extraction. The flexibility of schema, consistency and indexing policies, along with the capacity for rapid scale-out, make distributed NoSQL (Not Only SQL) data warehousing a standard for social big data. Most prominent examples today include Mongo DB [Cho10], Elastic Search [KBHG14] and Microsoft Cosmos DB [RP18, Gua18].
- **Velocity:** The online social media related requests which create, modify or remove UGC subject of analysis demand new algorithms and methods to manage and analyse the online and streaming data. The rate of incoming requests observed in this project ranges from 2000 to 8000 per second. Failure to cope with such traffic exposes the analysis to data loss or privacy non-compliance. Scalability and reliability of ingestion pipelines, therefore, become of high importance. These goals are historically at odds with each other and motivate radical innovation in distributed computing [Bre15, MFGZ18] and inter-process communication [Lea16].
- **Variety:** The structure, quality and completeness of SBD varies between networks and within a single platform over time. More often than not, incoming social UGC is collected in a semi-structured form, with structured metadata and high-dimensional unstructured (e.g., visual) attachments. Collection of such information introduces additional normalisation responsibility for the ingestion pipelines, before storage at non-SQL destination suited for unstructured SBD, exemplified by [Gua18, RSD+17].
- **Value:** The process of extracting knowledge from large volume of social data is referred to as big data analytics. Value is the most important characteristic of a big-data-based application because it allows generating useful business information [BOJC16]. The process of extracting value

is highly contextual, i.e., dependant on the domain of knowledge pursued. The extensive computational workload here is focused primarily on minimising the signal-to-noise ratio, defined within the domain's context. On a record level, this commonly entails filtering, aggregation and conflation of records. On an attribute level, the tasks include dimensional reduction, parsing, feature extraction and other task-specific transformations of the otherwise noisy information. These common tasks are expensive at scale and highly parallelizable, and as such benefit greatly, from the Map-Reduce [DG10] paradigm in big data analytics, standardised originally by Hadoop [She16, BOJC16], accelerated by Apache Spark [MBY+15, BOJC16, GA15], and commoditised by Databricks [XDG+14].

- **Veracity:** Behind any information management practice lie the core doctrines of data quality, listed by [BOJC16] as data governance and metadata management, along with considerations of privacy and legal concerns. Within this project, veracity refers to the correctness and accuracy of the information, both collected and inferred. The additional responsibilities of social big data in this category are substantial. The European General Data Protection Regulation (GDPR and ISO/IEC 27001) in force since May 25th, 2018, demands dedicated processes to collect, filter and apply all relevant privacy requests. These, in turn, require a custom NoSQL partitioning, to minimise the impact on overall storage and analytics performance. Privacy regulation also makes black-box approaches (like DNNs) to extracting value more difficult to use in business, requiring the results to be explainable on-demand [HBPK17]. The ability to explain the predictions (and biases) learned from SBD to the customer, regulator or the analyst himself, becomes an important responsibility of SBD analytics with social and computational consequences [Mil19]. Homogeneous factions [Axe97] plaguing social networks, including "echo chambers" [Sun14] and "filter bubbles" [Par11] can bias the collected data and reduce the accuracy of the study, further motivating (my) extended time-frame sampling, on Twitter enabled by Historical PowerTrack [TSS17].

1.3.2 Machine Learning Systems

In this era of unprecedented technological development, machine learning (ML) is recognised as one of the most important application areas, with the adoption gaining momentum across almost all industries [LS20]. The ever increasing market demand fuels the development of powerful toolkits and frameworks [KMF+17, GA15]. From autonomous cars and adaptive email-filters to predictive cyber-security and natural language processing, ML systems outperform humans on specific tasks [MKS+13, SHM+16, HZRS16b, LWJ+20] and increas-

ingly often, guide processes of human decisions and understanding [CHJ+16, DVGK14, DVK17]. However, as explained by [SHG+15], it is dangerous to expect such wins to come for free, as it is common to incur massive ongoing maintenance costs in real-world ML systems. [SHG+15] offers a comprehensive guide to uncovering and mitigating the hidden debts accumulated by ML systems. Below outlines additional effort necessary to mitigate several ML-specific risk factors in delivering knowledge from social data at scale.

> "Only a small fraction of real-world ML systems is composed of the ML code. The required surrounding infrastructure is vast and complex." (Sculley, 2015)

- **Anticipating Change**: One of the things that makes an ML system so fascinating in this project, is that it interacts directly with the external world. In 2020 we are reminded again that this world is rarely stable. The background rate of change creates an ongoing maintenance cost in delivering knowledge from social big data. In contrast to one-off delivery of an ML model in a scientific report, delivery in an industrial setting of this project requires ongoing adaptation to production and consumption patterns changing individually. Failure to do so can lead to system outages and deteriorating model performance. Maintaining high predictive performance requires regular monitoring, retraining and redeployment of the system, to adapt to changing culture and dynamics of attention online. These tasks alone motivate important technological advancements by Microsoft Azure and Databricks [MG18, Jos20, XDG+14].
- **Abstraction Boundaries**: Traditional software engineering has proven that encapsulation and modular design help create maintainable code. Careful separation of concerns and compartmentalisation of running code helps limit the cost of development, testing and unforeseen consequences of a single modification. Until recently there was however a distinct lack of strong abstractions to support ML systems, as noted by [SHG+15, Zhe14]. Today technology like Docker [LD19] and Kubernetes [Bre15] lend highly relevant paradigms to ML systems, by introducing abstraction layers between the ML code, compute platform and all the consumers.
- **Dependency Management**: The exposure of an ML system to external change increases with each dependency taken. Recent pre-trained models used to enhance UGC feature representation in popularity modelling are often subject to ongoing development. The model consuming it is likely to fit an older version and becomes exposed to immediate ramifications from a silent update of the pre-trained DNNs. ML systems depending on legacy data features (e.g., Twitter author's timezone, or GEO location) are subject to an immediate outage, once the information is no longer available.

> Careful feature reduction ahead of operationalisation is therefore critical and should also include a cost-to-benefit analysis of the more expensive dependencies (e.g., pre-trained DNN's for visual embeddings), to secure the scalability of the system. Upfront adoption of abstraction boundaries helps express the invariants and logical consistency of the information inputs and outputs from all involved components [SHG+15, Fow18]. Careful isolation and definition of components with their dependencies, enables testing, configuration and deployment, orchestrated by a new breed of ML DevOps services [MG18].

[SHG+15] delivers a compelling argument for thinking holistically about engineering (e.g., data collection) and research (e.g., feature extraction) concerns when designing ML systems. He attributes much of the ongoing cost of innovation to the still common and hard separation between data science and engineering roles. A hybrid approach where research and engineering concerns are considered in close alignment within the same team can help reduce the accumulation of technical debt [SHG+15] while rapidly accelerating ML innovation [SNP12]. I believe this approach describes a fundamental opportunity of industry-academic partnerships today, and I intend to illustrate it within this one.

1.3.3 Explainable AI

As the adoption of AI and ML systems gains momentum across almost all industries [LS20], [DVK17] highlights a surge of interest in systems optimised for another criterion besides the expected task performance, namely interpretability. In the context of ML systems, [DVK17] define interpretability as the ability to explain or to present in understandable terms to a human. However, a formal definition of an explanation remains elusive. In the field of psychology [Lom06] notes that questions such as what constitutes an explanation, what makes some explanations better than others, or how explanations are generated are just beginning to be addressed. Researchers around the world are proposing methods for interpreting complex ML models [Sam19, SWM17] with new urgency. The European General Data Protection Regulation (GDPR and ISO/IEC 27001) in force since May 25th, 2018 requires algorithms that make decisions based on user-level information, to provide an explanation [DVK17, GF17]. The motivation for explainability extends much further. [DVK17] argue that interpretability can assist in qualitatively ascertaining whether other desirable properties are met by the ML system, such as: fairness [BY14, HPS16, CJS18], safety [Ott13, AOS+16], avoiding technical debt [SHG+15], privacy, reliability and robustness among others. However, in many cases, formal definitions

remain elusive. The current state-of-the-art in interpreting complex models involves finding simplified models that provide explanations. Today predictions of any classifier or regressor are explained by approximating them locally with an interpretable model. Most popular of them today include: LIME [RSG16], DeepLIFT [SGK17], Layer-Wise Relevance Propagation [BBM+16], Shapley regression values [Lip06] and quantitative input influence [DSZ16]. Each of them presents a unique opportunity for scientific understanding. Applicability, accuracy and computational cost of these methods vary with the choice of ML technique to explain.

1.4 Conclusions

The purpose of this review was to summarise the trends and challenges in studying the dynamics of human attention, in the times where so much of it is aggregated and commoditised by the online social networks. As the average span of our collective attention shortens, our society's capacity to assimilate new knowledge and reject low quality influence diminishes [WFVM12, QFS+17]. In the world where the influence of individuals and organisations is built on attention capital, the struggle for attention replaces the struggle for material goods. Attempts at improving this struggle for both the producers and consumers of attention, while alleviating any negative consequences on our wellbeing or the democratic process itself, motivate abundant activity in the field of computational social science. Online social networks today measure attention received via an increasingly diverse set of engagement metrics, however extracting and delivering knowledge from social big data have proven to be extremely difficult, even before GDPR. Data protection regulations further complicate scientific and industrial inquiries, with the enforcement the privacy rights: to explanation and to be forgotten.

The era of big data offers significant advancements for data collection, analysis and delivery of machine learning systems, creating new opportunities for researchers to achieve high relevance and impact. However, much of the recent work is still plagued by anecdotal evidence from small, short time frame samples or black-box approaches challenging to interpret for the analyst, let alone the market in an industrial setting. It becomes clear that many obstacles of progress in the area stem from overly separated research and engineering roles undoubtedly inspired and reinforced by the old divide between industry and the academia. These divides are perhaps the root cause for much of the relevant data, experience and technology siloed in separation. It is argued that in order to move computational social science to the next level, the scientific community must meet both the challenges to deep understanding and re-usable

computational technology. Consequently, scientific and industrial pioneers in this area are relying on both social and computer science to impact the new frontier. This work aims to advance the understanding of the dynamics of collective attention. For it to be successful, a true merger of data science and data engineering appear necessary. The industrial research grant, awarded by the Danish Innovation Fund positions this project above the old divides for the period of three years. In the light of prior work both scientific and industrial, the immediate opportunities of a holistic approach become clear:

- to achieve controlled and meaningful observations of real-world phenomena from frequent, scalable experiments, via a careful fusion of data science and big data engineering advancements
- to deliver the new models of understanding to the academic community and Microsoft customers worldwide, in order to inform and where possible alleviate the struggle for attention at scale

The following Chapter 2 offers a detailed report on three studies modelling social engagement at scale, conducted after a careful fusion of recent technology, proposed in Chapter 3 in the form of a new data collection, analysis and operationalisation framework.

Chapter 2

Social Data Science

> "The role of the social and professional networks in the spread and acceptance of innovations, knowledge, business practices, products, behavior, rumors, and memes, is a much-studied problem in social sciences, marketing and economics. Online environments like Twitter, offer an unprecedented opportunity to track such phenomena."
>
> (Albert-László Barabási, 2016)

2.1 Introduction

This chapter summarizes the project's scientific results in modelling engagement in online social networks. It is organized into three parts. In [KL19], we revisit user-generated content (UGC) virality prediction on Twitter. I propose a new framework, which I describe in Chapter 3, for the rapid acquisition of large-scale datasets, high accuracy supervisory signal and multi-language sentiment predictions. We then employ a recent gradient boosting framework, to explore the limits of content virality prediction on Twitter, based on features available early (at the time of posting), while respecting every privacy request applicable. In [KH20], we take issue with the fact, that social influence today is still built on attention capital. In this study we examine and consolidate a diverse set of

content engagement metrics available from Twitter, in search of a better predictor. The correlations discovered allow us to propose a one-dimensional engagement signal, a more holistic alternative to virality, for influence maximization frameworks. The section provides a detailed explanation of the models, with real-world illustration of performance improvement. In the third and final study, we explore the transferability of the used methods across platforms [RKH20]. I use the data collection framework proposed in Chapter 3 to gather content and engagement metrics from Instagram. We then build and use a novel feature extraction framework, to compute a rich, multi-modal content representation of Instagram posts. We present a new strong baseline for popularity prediction on Instagram which is both robust and efficient to compute. The final ablation study quantifies the impact of each modality to achieving strong [Coh88] ranking performance. Each contribution aims to advance online social engagement prediction in a scalable and explainable way [KL19, KH20, RKH20].

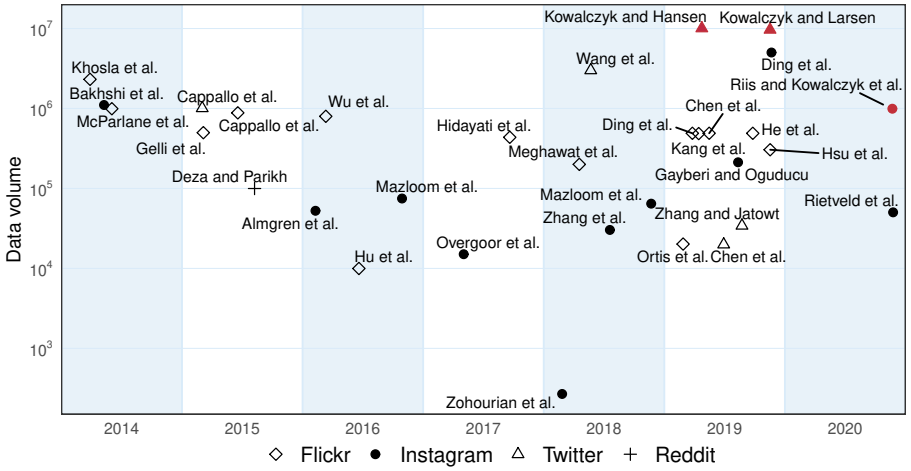


Figure 2.1: The volume of data used in prior work modeling the dynamics of online social engagement varies. This Ph.D. project (red marks) is among the largest studies on social media to date.

2.2 Virality Prediction on Twitter

In [KL19] we seek to maximize virality ranking performance under scalability and explainability constraints. We follow [WBF18] to approach the problem as Poisson regression and [HAN+11] to consider the tweet sentiment in prediction.

However, in the contrast to prior work, we rely only on features available immediately after posting, and avoid black-box approaches (e.g., deep learning), to ensure model's scalability and explainability in production. The contributions of this study are summarized as follows:

- A novel framework for rapid collection and analysis of Twitter UGC at scale, summarized by Chapter 3
- A novel feature representation of Twitter UGC, based entirely on features available early
- A new model for predicting virality on Twitter, achieving state-of-the-art ranking performance, validated on multiple benchmark datasets
- Feature importance analysis of the proposed virality model

2.2.1 Data collection

I use the new framework described in Chapter 3, to collect multiple training and benchmark datasets, retroactively from the Twitter archive. Table 2.1 offers a summary of all the tweets analyzed in this study. 85% of these tweets have never been retweeted and illustrate the asymmetry of attention in Twitter. I have computed the sentiment score and collected the count of retweets ever registered for each of the tweets collected.

Dataset	**MBI**	**T2015**	**T2016**	**T16-BIO**	**T2017-BIO**
Introduced	[CMS15]	[WBF18]	[WBF18]	[KL19]	[KL19]
Data from	2013.02	2015.11	2016.10	2015.06	2017.01
Date to	2013.03	2016.04	2015.12	2017.06	2018-02
Months	2	6	3	12	14
Language	English	English	English	Multi (18x)	Multi (18x)
w/images only	TRUE	TRUE	TRUE	FALSE	FALSE
Tweets Total	2,724,764	9,025,826	8,469,016	27,032,417	19,850,448
Unique (acquired)	1,319,288	2,804,153	2,736,600	14,788,552	9,719,264
Never retweeted	1,042,411	2,106,475	2,088,377	12,809,021	8,774,009

Table 2.1: Datasets acquired. Based on [KL19].

2.2.1.1 Benchmark datasets

I collect three benchmark datasets MBI, T2015 and T2016 (with a total of 6,860,041 unique tweets) to enable comparative analysis with the work of [MRR+16,

MMJ14, KDH14, CMS15, WBF18]. During collection I have used the same filters (e.g., date, language or presence of an image attachment) as reported in prior work, yet retrieving higher volume. The datasets are then split 70/20/10 for training/test/validation, similar to [WBF18, CMS15].

2.2.1.2 Extended time-frame datasets

The framework I propose in Chapter 3, facilitates analysis at a significantly larger scale than above. In an attempt to maximize the model's generalizability over languages and cultures, I have collected another 24 million unique tweets. These were carefully sampled across an extended time-frame of 20 months, and 18 languages supported by the company's sentiment analysis in 2018 [Mic17].

2.2.2 Feature representation

I use the feature extraction framework, proposed in Chapter 3, to compute feature vectors for every Tweet collected. Multiple features have been selected and pre-processed from the raw UGC. They aim to represent content, the author, the time of posting and they way (sentiment). Table 2.2 describes the feature vector and their Pearson correlation coefficient with the logarithm of retweet count in the T2017-BIO dataset. Some authors receive more attention than others despite low activity. The two author ratio features are computed in an attempt to isolate them. Number of attachments is an aggregated count of hashtags, mentions, URLs, images, symbols and videos. Finally, I follow [COM13] to log-transform the power-law distributed author features (e.g. author's favorite and listed counts). With scalability in mind, only features available at the time of posting or immediately after are considered. I do not consider the early performance of the tweet nor image-based features at this point.
I acquire the retweet count ever registered for above tweets via Twitter's Engagement API (Chapter 3). The counts are used as ground truth after log-transformation, due to power-law distribution [COM13] to stabilize variance:

$$r = \ln(\epsilon_{retweets} + 1) \tag{2.1}$$

2.2.3 Gradient Boosted Poisson Regression

In [KL19], I consider the problem of predicting the scale of retweet cascade for any given tweet, using only features available immediately after posting. The

Modality	Feature	Type	Pearson
(A) Author	followersCount	ordinal	0.205920
	friendsCount	ordinal	0.082779
	accountAgeDays	ordinal	0.020379
	statusesCount	ordinal	-0.001455
	actorFavoritesCount	ordinal	0.029914
	actorListedCount	ordinal	0.221067
	actorVerified	categorical	0.202722
(C) Content	attachmentCount	ordinal	0.085333
	mentionCount	ordinal	-0.006590
	hashtagsCount	ordinal	0.104335
	mediaCount	ordinal	0.147623
	urlCount	ordinal	0.082549
	isQuote	categorical	0.061915
(L) Language	languageIndex	categorical	0.005199
	sentimentValue	continuous	0.059863
(T) Temporal	postedHour	ordinal	0.016639
	postedDay	ordinal	-0.000963
	postedMonth	ordinal	-0.004129
	postedDayTime	categorical	0.016639
	postedWeekDay	categorical	-0.001002

Table 2.2: Twitter UGC representation summary: features extracted to represent the tweet's author (A) content (C) language (L) and the time of posting (T). Only weak linear correlation with virality observed.

author features are used together with the content, language, and temporal to predict the number of future retweets. I follow [WBF18] to assume the future retweet count r follows Poisson distribution:

$$P(R = r \mid \lambda) = \frac{e^{-\lambda}\lambda^{-r}}{r!} \tag{2.2}$$

where the latent variable $\lambda \in R^+$ defines the mean and variance of the distribution. Fitting such distribution involves maximizing the Poisson log-likelihood, given a collection of N training tuples of tweets t_i and their retweet counts r_i:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum [r_i \ln \lambda(t_i) + \lambda(t_i)] \tag{2.3}$$

where θ contains all parameters of my proposed model.

2.2.4 Gradient Boosted Regression Trees

Gradient Boosted Regression Trees (GBRT) is a tree ensemble algorithm which builds one regression tree at a time by fitting the residual of the trees that preceded it. Considering the twice-differentiable loss function assumed in this study, and denoted as:

$$L_{\text{Poisson}}(r, t) = r \ln \lambda(t) + \lambda(t) \tag{2.4}$$

GBRT minimizes the function (regularization omitted for simplicity):

$$L = \sum_{i=1}^{N} L_{\text{Poisson}}(r_i, F(t_i)) \tag{2.5}$$

with a function estimation F(t) represented in an additive form:

$$F(t) = \sum_{m=1}^{T} f_m(t) \tag{2.6}$$

where each $f_m(t)$ is a regression tree and T is the number of trees. GBRT learns these regression trees in an incremental way: at m-stage, fixing the previous $m-1$ trees when learning the m-th trees [BFSO84, KL19]. To construct the m-th tree, GBRT minimizes the following loss:

$$L_m = \sum_{t=1}^{N} L_{\text{Poisson}}(r_i, F_{m-1}(t_i) + f_m(t_i)) \tag{2.7}$$

where $F_{m-1}(t) = \sum_{k}^{m-1} f_k(t)$. Eq. 2.8 solves the optimization problem in (2.7) via Taylor expansion:

$$\begin{aligned} L_m \approx \bar{L}_m = \sum_{i=0}^{N} [& L_{\text{Poisson}}(r_i, F_{m-1}(t_i)) \\ & + \nabla_i f_m(t_i) + \frac{\nabla_i^2}{2} f_m^2(t_i)] \end{aligned} \tag{2.8}$$

The gradient and Hessian in Eq. 2.8 are defined as:

$$\begin{aligned} \nabla_i &= \frac{\partial L_{\text{Poisson}}(r_i, F(t_i))}{\partial F(t_i)} \mid_{F(t_i)=F_{m-1}(t_i)} \\ \nabla_i^2 &= \frac{\partial^2 L_{\text{Poisson}}(r_i, F(t_i))}{\partial^2 F(t_i)} \mid_{F(t_i)=F_{m-1}(t_i)} \end{aligned} \tag{2.9}$$

I train the GBRT by minimizing $\bar{L}_m$ which is equivalent to:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \frac{\nabla_i^2}{2} (f_m(t_i) + \frac{\nabla_i}{\nabla_i^2})^2 \tag{2.10}$$

The Poisson loss function is vulnerable to over-dispersion and power-law distribution, characterizing the retweet count. Only 16% of the collected tweets have ever received a retweet (Table 2.1). In cases where the Hessian is nearly zero, Eq. 2.10 approaches infinity. To prevent the gradient explosion and safeguard the optimization process, I cap each tree's weight estimation at 1.5.

2.2.5 Predicting Virality

In this study, the predictive analysis of UGC virality on Twitter, begins with experiments on the benchmark datasets, to evaluate the proposed feature representation against recent state-of-the-art methods. The ablation study follows, to examine the limits of the approach on the extended time-frame datasets. There, I pursue a scalable model, that could generalize across topics (months) and cultures (languages).

2.2.5.1 Evaluation metrics

In order to measure the models' ability to rank the content by expected virality, we compute the Spearman Rho ranking coefficient. Interpretation of this metric is domain specific, with guidelines for social/behavioral sciences proposed by [Coh88]. SciPy version 1.4.0 is used to ensure tie handling. I did not find this concern expressed in work prior to [KL19]. I choose relative (R^2) and absolute (RMSE) measures of fit for tuning. RMSE penalizes large error higher (i.e. when underestimating highly viral content or vice-versa) in the contrast to MAPE. We have agreed to disregard MAPE relative to above when fitting an asymmetric, zero-inflated distribution of a dependent variable (like retweet count). MAPE is undefined for the majority of examples (Table 1), which never receive a retweet and penalizes errors for least retweeted higher [KL19].

2.2.5.2 Method validation experiments

For a fair comparison with previous SOTA, I use Poisson regression on the joint author, content and temporal features (ACT), before including sentiment (ACTL). Table 2.3 offers the results of modeling on benchmark datasets. The GBRT achieves substantially higher ranking performance, compared to other content-based methods, already before considering image and propagation modalities. The p-value for all reported Spearman results is $p < 0.001$.

Method	**SpearmanR**			**MAPE**		
	MBI	**T2015**	**T2016**	**MBI**	**T2015**	**T2016**
[MMJ14]†	0.188	0.269	0.257	0.093	0.121	0.137
[KDH14]†	0.185	0.273	0.254	0.097	0.103	0.124
[CMS15]†	0.189	0.265	0.258	0.089	0.095	0.119
[MRR+16]†	0.190	0.287	0.262	0.073	0.097	0.117
[WBF18]†	0.229	0.358	0.350	0.057	0.084	0.103
ACT	**0.322**	**0.498**	**0.503**	0.247	0.266	0.256
ACTL	**0.323**	**0.499**	**0.504**	0.247	0.266	0.255
	R^2			**RMSE**		
	MBI	**T2015**	**T2016**	**MBI**	**T2015**	**T2016**
ACT	**0.303**	**0.417**	**0.391**	**0.444**	**0.553**	**0.555**

Table 2.3: Based on [KL19]. Method performance on benchmark datasets. † independent evaluation by [WBF18].

The proposed feature representation, including sentiment score and a high accuracy ground-truth, allows me in 2018 to outperform the state-of-the-art by more than 37% on multiple benchmark datasets.

2.2.5.3 Multi-lingual, extended time-frame ablation study

I apply the same feature extraction and GBRT optimization, to the newly-acquired T2017-BIO dataset, in order to maximize a model's performance across languages and time. Every tweet in the test dataset is again described by the content (C), author (A), language (L) and temporal (T) features. Table 2.4 summarizes the contribution of these dimensions individually and in combination. The baseline model is trained on a single feature, prevalent in prior work: the number of followers, to be notified at the time of posting.

2.2.6 Feature importance and discussion

The relative insignificance of the temporal features (T) indicates low correlation between the time of posting and the content virality. This would challenge the common intuition to post at the common intuition, to post at the right time, helps propagating the content. I also observe that the content-based features alone outperform (Spearman 0.211) the followers baseline in virality ranking. On Twitter, how many people follow you appears less important than what you have to say.

Features	SpearmanR	R^2	RMSE	MAPE
A	0.310	0.317	0.359	0.133
C	**0.211**	0.055	0.422	0.160
T	0.062	0.001	0.432	0.171
L	0.164	0.017	0.430	0.167
AC	0.356	**0.396**	0.337	0.121
AT	0.311	0.316	0.359	0.132
AL	0.324	0.320	0.358	0.130
CT	0.220	0.059	0.421	0.159
CL	0.269	0.076	0.417	0.154
TL	0.170	0.019	0.430	0.166
ATL	0.324	0.320	0.358	0.130
ACT	0.357	0.395	0.338	0.120
ACL	0.369	0.399	0.336	0.119
ACTL	**0.369**	**0.402**	**0.336**	**0.118**
Baseline	**0.180**	**0.091**	**0.414**	**0.160**

Table 2.4: Ablation results: quantitative evaluation of 'A': actor, 'C': content, 'T': temporal, and 'L': language features in predicting virality on Twitter. Based on [KL19].

The use of a tree ensemble method, like the gradient boosting machine in this study, provides an opportunity for model explanation via feature importance. A classic approach based on gain, has been introduced by [BFSO84]. Gain is the total reduction of loss or impurity contributed by all splits based on a given feature. Another common approach is simply to count how many times a fea-

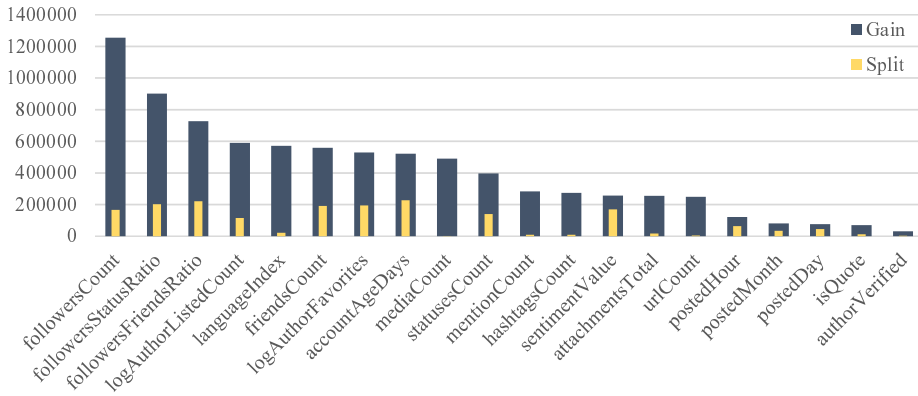


Figure 2.2: Feature level importance, as measured by the total purity gain or the number of splits a feature participated in. Based on [KL19].

ture is used when growing a tree. Feature splits are arguably most intuitive, in representing a feature's importance [LEL18]. Figure 2.2 illustrates the importance of features used by the ACTL model, trained on the extended time-frame dataset T2017-BIO.

The size of the audience immediately exposed to the tweet (followersCount) is the highest contributor to the model performance. The number of followers gained per tweet and per friend, or the number of times an author has been listed by others, are also consistently predictive of virality. Figure 2.2 has inspired many hypothesis for further study. The number of friends could inform the diversity of content the author is exposed to. The count of tweets favorited over time (e.g., the age of account) can also be indicative of the author's overall engagement with the platform. With the author's social influence defined as the ability to spread information in the network [PAP$^+$13], could the diversity of content actively consumed over time maximize authors I propose this hypothesis for computational social science [KL19].

2.3 Beyond Virality: Compound Engagement

In [KH20], we question virality as the best metric for maximizing influence in online social networks. Much of prior work on influence analysis is focused on homogeneous information networks and attributes the greatest influence on authors who trigger the largest diffusion cascades [Fra19]. When the author's influence is modelled as the ability to maximize the expected spread of information in the network, the most desirable user-generated content is the one propagated furthest, in Twitter measured by the number of retweets [PAP+13, EMR+19, KH20]. Virality, however, does not capture the average individual attention received. Retweet action does not inform, e.g., if the actor has actually read the content, let alone consider the source or whether that effort was left to the followers [KH20]. Meanwhile, the abundance of information to which we are exposed through online social networks is exceeding our capacity to consume it, let alone in a critical way [WFVM12, QFS+17]. [LSMHL19] shows that the competition for our attention is growing, causing individual topics to receive even shorter intervals of collective attention.
Since the conception of influence maximization frameworks, multiple other engagement metrics became available. In [KH20] we examine and consolidate a diverse set of content engagement metrics to propose statistical evidence of a new one-dimensional engagement signal. We next show the relevance of the signal for understanding engagement in multiple datasets, and prove it is more predictable than the individual influence metrics (e.g., diffusion size measured by retweet count) assumed in prior work. The contributions of [KH20] are summarized as follows:

- Advanced feature representation of UGC on Twitter, one of the first to consider increasingly popular 'quote tweets', validated on two real-world datasets
- New state-of-the-art in virality prediction on Twitter
- Two new engagement models: response and popularity, delivering strong ranking performance based on user generated content (UGC) features available at the time of posting
- Evidence of one-dimensional engagement signal on Twitter
- A new compound engagement formula, capturing over 75% of variance in the available discrete signals
- A new more holistic, compound engagement model, first to explain half of the variance with content features available early, and to offer strong ranking performance simultaneously

2.3.1 Data collection

I use the framework proposed in Chapter 3 to collect training and validation datasets described in Table 2.5. Retroactive filtering of the Twitter archive enables me to closely reproduce the datasets used in prior work. The framework also facilitates a near-uniform sampling across extended time-frames, in order to increase the size of the population represented by the sample, as motivated by [KJKW18]. To enhance UGC representation and ensure fair comparison with the earlier results, I reuse the sentiment predictions from [KL19, Mic17]. The summary of one benchmark and two extended time-frame datasets acquired for this study is offered by Table 2.5 and motivated below:

Dataset	**T2016-IMG**	**T2017-ML**	**T2018-ML**
introduced	[WBF18]	[KL19]	[KH20]
w/image only	True	False	False
languages	**English**	**18**	**all**
months total	**3**	**14**	**12**
month from	2016.10	2017.01	2018.01
unique tweets	2,848,892	9,719,264	29,883,324
quoting	421,175	583,514	2,647,072
retweets total	5,929,850	11,361,699	42,919,158
replies total	717,644	3,576,976	12,414,907
favorites total	12,665,657	29,138,707	134,523,998
no engagement	1,547,829	5,689,501	14,813,772

Table 2.5: Datasets acquired. Based on [KH20].

- **T2016-IMG** to evaluate both a new feature representation and method in comparison with the work of [MRR$^+$16, MMJ14, KDH14, CMS15, WBF18, KL19]. The dataset again matches the same filters, as applied before.

- **T2017-ML** to evaluate the generalizability of the new method and representation in comparison with the previous study [KL19]. This dataset represents a near-uniform sample of all Twitter UGC posted over 14 months. It is filtered to 18 languages supported by our sentiment analysis [Mic17].

- **T2018-ML** to evaluate the generalizability of the new compound engagement model across years. This dataset represents a near-uniform sample of the entire Twitter 2018 volume, in all known languages. Due to time constraints, I use this dataset in unsupervised experiments only.

2.3.1.1 Discrete Engagement Signals

I use the new framework also to retrieve the number of retweets, replies and favorites for every tweet in this study. These represent individual engagement ever registered for a tweet (even if removed later). The use of Twitter Engagement API, as proposed in Chapter 3, ensures 100% accuracy of the engagement signal:

$$\epsilon = [\epsilon_{\text{retweets}}, \epsilon_{\text{replies}}, \epsilon_{\text{favorites}}]^T. \tag{2.11}$$

2.3.2 Extending feature representation

[GWDC16] demonstrate the impact of "quote retweets" on political discourse and its diffusion. The feature to quote another tweet has been introduced by Twitter in 2015, yet very few studies consider them in UGC feature representation. Table 2.5 shows that over 3.5 million tweets analyzed in this study quote another. I extend the feature selection proposed in [KL19], to represent them. Table 2.6 offers a summary of the new feature representation. The additional 14 unique features extracted for quoting RTs are shown in bold. To stabilize variance, I log-transform the highly skewed (e.g., count of followers, friends, statuses or the number of times the author has been listed). To ensure scalability in production, only the information available at the time of posting is considered.

2.3.3 Compounding Engagement

In this section I examine the multi-dimensional content engagement vectors ϵ acquired for the extended time-frame datasets. I use Parallel Analysis to look for potential correlations that could enable reducing the dimensionality of UGC engagement on Twitter.

2.3.3.1 Principal Engagement Component

Recent work on engagement modeling, defines any response as a sign of engagement, effectively reducing the multivariate response to a one-dimensional signal [LHN18, KH20]. However, to the best of our knowledge, the complexity of the engagement signal has not been explored more formally. In [KH20] we hypothesize that the population response signals, i.e., the dimensions of the of vector $\mathbf{e}$, are highly correlated and proceed to test the effective dimension of the space populated by the vectors using Parallel Analysis (PA) [Hor65, JH11].

Feature	Treatment	Skewness	Quoted†
followers count	ordinal	0.212	**True**
friends count	ordinal	-0.321	**True**
account age (days)	ordinal	0.203	**True**
statuses count	ordinal	-0.665	**True**
actor favorites count	ordinal	-1.023	**True**
actor listed count	ordinal	0.687	**True**
actor verified	categorical	-	**True**
body length	ordinal	-1.426	**True**
mention count	ordinal	3.820	**True**
hashtag count	ordinal	5.808	**True**
media count	ordinal	3.203	**True**
URL count	ordinal	1.449	**True**
language code	categorical	-	**True**
sentiment value	continuous	-0.014	False
posted hour	ordinal	-0.058	False
posted day	ordinal	0.021	False
posted month	ordinal	0.210	False
retweet count	label	6.091	n/a
reply count	label	2.330	n/a
favorite count	label	3.122	**True**

Table 2.6: Based on [KH20]. Summary of the feature representation for UGC on Twitter,
now extended to also represent the tweets quoting another.
† if True, additional feature is extracted from the quoted tweet.

In PA, principal component analysis (PCA) of the measured signals is compared with the distribution of the principal components of null data, which is obtained by permutation under a (null) hypothesis. The null hypothesis assumes no correlation between the individual engagement signals. Consistent with this hypothesis, we can permute the sequence of the signals for each observation separately. In particular, we calculate the upper 95%quantile for the distribution of the eigenvalues in the permuted dataset. Eigenvalues of the original unpermuted dataset that reject the null hypothesis are considered 'signal'. Principal components are computed on the engagement signals after variance stabilization:

$$\varepsilon = \ln(\epsilon + 1), \tag{2.12}$$

similar to treatment of the retweet count by [COM13, KL19].

2.3.3.2 Projection on the engagement component

Hypothesizing the one-dimensional engagement signal, I compute the value by projecting the transformed $D = 3$ dimensional data on the first principal component:

$$\mathrm{E}_1 = \sum_{\mathrm{i}=1}^{\mathrm{D}} \mathrm{w}_\mathrm{i} \left(\ln(\epsilon_\mathrm{i} + 1) - \mu_\mathrm{i} \right), \tag{2.13}$$

here $\mu_i = \frac{1}{N} \sum_{n=1}^{N} \varepsilon_{i,n}$ is the i'th component of the D-dimensional mean vector for a sample of size N, while w_i is the i'th component of the first principal component, computed on the same sample.

2.3.3.3 Evidence for a one-dimensional engagement signal

We exercise Parallel Analysis to compute the principal components and the variance of their associated projections for the log-transformed data and for an additional $Q = 100$ permutations of the data which assumes no correlation (null hypothesis). The one-sided upper 95% quantile is computed from the permuted samples. Figure 2.3 shows variances of the un-permuted signals and the 95% quantiles for the three eigenvalues of the permuted data. Very similar results

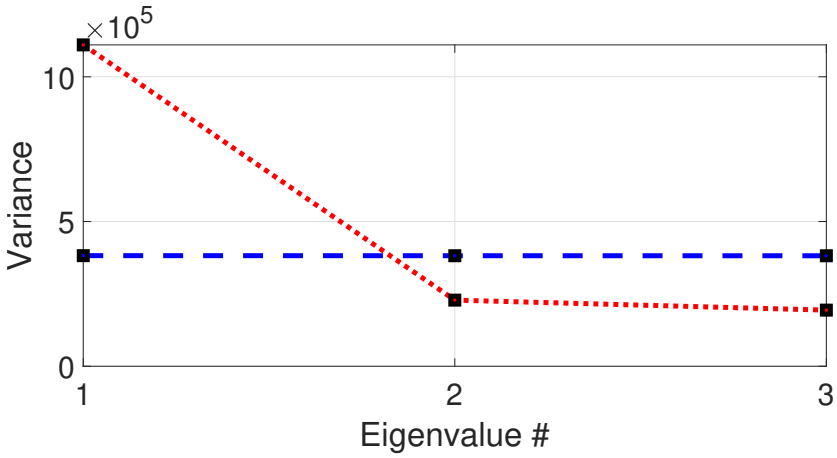


Figure 2.3: Evidence for a one-dimensional engagement signal in T2017-ML dataset. Parallel Analyses of the engagement signals shows only the first component ('1'- red dotted line) exceeds the 95% quantile of the corresponding eigenvalue in the null hypothesis (blue line). Based on [KH20].

are obtained for the T2018-ML dataset.

2.3.3.4 The engagement signal

We perform principal component analysis of the three datasets to find the components capturing the most variance in the discrete engagement signals. Table 2.7 offers the mean vectors and projections. The variance explained by the first components is $83\%, 72\%, 77\%$ for the analysis of T2016-IMG, T2017-ML and T2018-ML respectively.

	retweets		**replies**		**favorites**	
	w_1	μ_1	w_2	μ_2	w_3	μ_3
T2017-ML	0.451	0.049	0.145	0.082	0.880	0.148
T2018-ML	0.450	0.066	0.188	0.080	0.872	0.205

Table 2.7: Based on [KH20]. First principal components of the engagement signals present in the extended time-frame datasets. The components are used to compute the one-dimensional compound engagement (see Equation 2.13)

2.3.4 Predicting Compound Engagement

In the supervised experiments, I first evaluate the feature representation against previous state-of-the-art methods, by modelling the individual influence metrics (e.g., size of diffusion), and the compound engagement on the benchmark dataset T2016-IMG [KH20]. I then proceed to evaluate the generalizability of the new methods across topics and cultures on the multilingual extended-timeframe dataset T2017-ML. Both datasets are split into 70% training, 20% test and 10% validation sets.

2.3.4.1 Gradient Boosted RMSE Regression

I consider the problem of predicting audience engagement for a given tweet based on features available immediately after its delivery (Table 3). Assuming the extended feature representation, proposed in section 1.3.2, I exercise gradient boosted regression to predict the number of retweets (i.e., the size of diffusion cascade), number of likes, replies and finally, the proposed compound

engagement signal. I use the tree ensemble algorithm GBRT, described in section 1.2.4. In this study I choose RMSE as the twice-differentiable loss function for the training process:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} L_{\mathrm{SE}}(\hat{\varepsilon}_i(\theta), \varepsilon_i), \tag{2.14}$$

where θ contains all parameters of the proposed model, N is the number of examples, and L_{SE} is the squared error of an individual prediction,

$$L_{\mathrm{SE}}(\varepsilon, \hat{\varepsilon}) = (\varepsilon - \hat{\varepsilon})^2. \tag{2.15}$$

The choice of RMSE over Poisson objective used before comes with a trade-off between predictive and training performance. In a few experiments not reported here, I have found that the RMSE objective leads to c.a. 2% lower ranking performance, however at a dramatic increase in training speed. RMSE loss function no longer requires each tree's weight to be capped, to prevent gradient explosion. Following [COM13, KL19] I stabilize variance of all individual engagement signals via log-transformation (see Equation 2.12).

2.3.4.2 Evaluation metrics

We compute the Spearman ρ ranking coefficient to measure each model's ability to rank the content depending on the definition of engagement [KH20]. We compute the relative measure of fit R^2 to compare the variance explained by the compound engagement and the individual engagement models. I choose the absolute measure of fit (RMSE) an objective of optimization, to penalize large errors higher and accelerate the training process, relative to Poisson objective.

2.3.4.3 Validating extended representation

First round of our supervised experiments focus on evaluating the extended UGC feature representation proposed in 1.3.2 and the GBRT approach against previous state-of-the-art methods, in experiments organized by the definition of engagement. I begin with modeling the established metrics, like the size of diffusion cascade (i.e., retweet count), response (i.e., number of replies) and popularity (i.e., number of favorites/likes), then proceed to modeling the compound engagement.
Table 2.8 shows the GBRT performance with RMSE objective and new feature representation, depending on the prediction target. The extended UGC feature

representation did not provide a significant boost over [KL19]. This is not surprising, considering the visual modality dominates the T2016-IMG dataset, as considered by [WBF18]. The analysis of visual features is out-of-scope for this study. The model did, however, match the performance of [KL19] in virality ranking, and achieves strong [Coh88] performance already before including image features. When applied to compound engagement prediction, it sets a new benchmark for content engagement ranking at $\rho = 0.680$.

2.3.4.4 Multi-lingual extended time-frame study

In the second round of supervised experiments we explore the scalability and generalizability of the approach across topics and cultures [KH20]. Table 2.4 shows the performance of the GBRT depending on the definition of engagement, on the multilingual extended time-frame dataset T2017-ML. Predicting the number of retweets with the new feature representation outperforms [KL19], offering new state-of-the-art in virality ranking. The response and popularity models each achieve strong ranking performance on T2017-ML [Coh88, KH20]. The compound engagement model again shows an increase in ranking performance as compared to the individual engagement models, and sets a new benchmark for engagement variance explained at $R^2 = 0.507$.

The p-value for all reported ρ results is $p < 0.001$. Each result is an average across folds in 3-fold cross-validation. I am using SciPy version 1.3.1 to ensure ρ tie handling. Interpretations of R^2 and Spearman ρ are domain-specific, with guidelines for social and behavioral sciences proposed by [Coh88]. Considering the exceptional amount of external confounders affecting predictive analysis in this domain, a model to achieve 0.5 is considered strong.

Method	R^2	ρ	RMSE
[MMJ14]†	-	0.257	-
[KDH14]†	-	0.254	-
[CMS15]†	-	0.258	-
[MRR+16]†	-	0.262	-
[WBF18]†	-	0.350	-
[KL19]	0.391	0.504	0.555
virality (retweets)	0.393	0.504	0.554
response (replies)	0.239	0.384	0.290
popularity (favorites)	0.500	0.656	0.665
engagement (compound)	**0.501**	**0.680**	**0.341**

Table 2.8: Based on [KH20]. Method evaluation against previous SOTA, on the benchmark T2016-IMG dataset
† independent evaluation by [WBF18]

Method	R^2	ρ	RMSE
virality [KL19]	0.402	0.369	0.336
virality (retweets)	0.425	0.371	0.329
response (replies)	0.302	0.512	0.292
popularity (favorites)	0.493	0.526	0.484
engagement (compound)	**0.507**	**0.529**	**0.228**

Table 2.9: Engagement prediction performance on the T2017-ML dataset
R^2, ρ: higher is better. RMSE: lower is better.
$SD < 0.001$ in 3-fold cross validation. Based on [KH20].

2.3.5 Real-world performance, illustrated

In the social media monitoring scenarios envisioned in this industrial project, the ability to predict the exact amount of engagement comes secondary to content ranking performance. In the times of limited attention, only a minority of content will receive it at scale. Deciding the order of this minority at any given time, is commonly approached as a ranking problem [Hea10, TAdAF14, TADF12, GIL12].

Table 2.10 offers some of the most prominent examples of Twitter UGC, ranked by the size of diffusion cascade (measured by the number of retweets). When the author's social influence is ranked by the ability to maximize the expected spread of information in the network [PAP+13, EMR+19, KL19, KH20], the number of likes (favorites) or replies are ignored.

Tweet (body)	Retweets	Replies	Favorites
"ZOZOTOWN新春セルが史上最速で取高100を先ほ(...)"	**4.5M**	357.4K	1.3M
"HELP ME PLEASE. A MAN NEEDS HIS NUGGS"	**3.47M**	37K	0.99M
"If only Bradley's arm was longer. Best photo ever. #oscars"	**3.21M**	215K	2.29M
"No one is born hating another person because of the color of his skin or his background or his religion..."	**1.61M**	69K	4.44M

Table 2.10: Four prominent tweets ranked by the influence predictor most popular in prior work: the size of diffusion triggered in the network, in Twitter measured by the number of retweets [KH20].

Table 2.11 illustrates ranking performance by the new compound engagement metric, in a striking contrast with the traditional diffusion-based approach. The tweet considered least influential in ranking 2.10, by the former US president Barack Obama (quoting Nelson Mandela), is now ranked first.

Tweet (body)	Engagement
"No one is born hating another person because of the color of his skin or his background or his religion..."	**9.283**
"If only Bradley's arm was longer. Best photo ever. #oscars"	**9.266**
"ZOZOTOWN新春セルが史上最速で取高100を先ほ(...)"	**9.158**
"HELP ME PLEASE. A MAN NEEDS HIS NUGGS"	**8.822**

Table 2.11: The four prominent tweets, ranked by the new compound engagement metric [KH20].

2.3.6 Explaining engagement, discussion

The commercial opportunity of the models proposed in this project, to inform real-world business decisions, requires the predictions to be explainable. The EU General Data Protection Regulation (GDPR) motivates AI researchers to propose various methods for interpreting complex models [VA19]. I apply some of the most popular and the most promising of these methods, in a novel attempt to explain online social engagement at scale.

2.3.6.1 Feature importance

I begin explaining social engagement models produced by this study with the classical approach introduced by [BFSO84]. Figure 2.4 offers a comparison of feature importance between all engagement models trained on the T2017-ML dataset [KH20]. The importance of each feature is calculated as total purity gain of splits which use the feature, averaged across 3-folds of cross-validation and then re-scaled to $[0, 1]$ for comparison across all models. The uncertainty for virality features does not exceed 6%. When predicting response (i.e., number of replies), the number of users mentioned by the tweet has the highest predictive value, in the contrast to the number of image attachments (i.e., media count) which has almost none. The count of followers, highly popular in prior work on virality prediction comes fourth in predicting the compound engagement. The average number of followers gained with each status (i.e., followerStatusRatio) or the number of times the author favorited other tweets (i.e., actorFavoritesCount) are far more predictive of compound engagement.Feature importance analysis based on the gain and split counts, offers only a model level heuristic towards explainability. Consistency of this approach is challenged by [LEL18]. Lundberg argues, that a model can change such that it relies more on a given feature, yet the importance estimate (gain or split count) assigned to that feature can decrease.

2.3.6.2 Consistent individualized feature attributions

The SHAP [LEL18] method computes Shapley values from coalitional game theory, by which the feature values of a single social post act as players in a coalition [LL17b, Mol19]. Shapley values tell us how to fairly distribute the 'payout' (= the engagement prediction) among the features. A player can be an individual feature value (e.g., 2 million followers), but also a group of feature values.

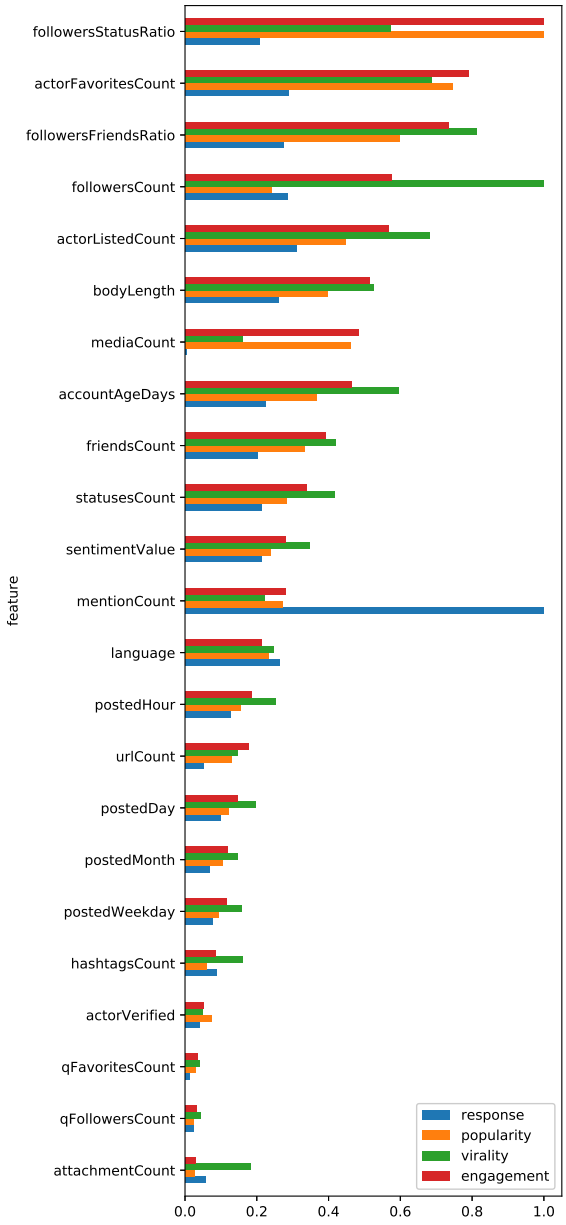


Figure 2.4: Relative feature importance depending on the definition of engagement, based on cumulative purity gain calculation (top 23 out of 31 features). Author features contribute the highest purity gains consistently across models, however contribution of other features, like the number of mentions or media attachments, shows significant variance across models. Based on [KH20].

Formally, a *coalitional game* is defined by a set N of n players, and a function v that maps subsets of players to the real numbers: $v : 2^N \to \mathbb{R}$, with $v(\emptyset) = 0$, where $\emptyset$ denotes the empty set (i.e., a coalition with all features missing). Function v called the worth of coalition S has the following meaning: if S is a coalition of players, then $v(S)$ describes the total expected sum of payoffs the members of S can obtain by cooperation, in this study defined by a specific engagement prediction for N. The Shapley value is one way to distribute the prediction (payout) to the features (players). It is a "fair" distribution in the sense that it is the only distribution with certain desirable properties (e.g., additivity explained further by [Mol19]). According to the Shapley value, the amount that the feature (player) j is given in a coalitional game (v, N) is:

$$\phi_j(v) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(v\left(S \cup \{x_j\}\right) - v(S)\right) \tag{2.16}$$

The sum extends over all subsets S of N not containing player j. When calculating for the feature *followersCount* in engagement prediction, the formula be can interpreted as follows. For each of the coalitions, I compute the predicted engagement with and without the feature value *followersCount* and take the difference to get the marginal contribution. The Shapley value is the (weighted) average of marginal contributions. I replace the feature values of features that are not in a coalition with random feature values from the test dataset to get the prediction from the engagement model. If I compute the Shapley values for all UGC feature values, I get the complete distribution of the prediction (minus the average) among the feature values. Thus SHAP specifies the explanation of a single engagement prediction as:

$$k(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2.17}$$

where k is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector of UGC input features, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature value attribution or simply the Shapley value, for the feature j of this UGC instance. To advance my understanding of social engagement at scale, I proceed with Shapley Analysis to compute explanations for predictions in the test set. Figures 1.4 to 1.7 offer a summary of Shapley Analysis performed for the extended time-frame engagement models trained in this study (see section 1.3.4.4).The SHAP summary plots are computed on a random sample of two thousand UGC instances from the test set, using the *shap* library from [LEL18]. The colour and position of the dots inform the impact on the prediction of the chosen feature's value. A red dot to the right informs that the high value of the particular feature contributed the most to the prediction. A blue dot to the left means that the low value of the feature reduced the engagement prediction. Violet dot in the middle means that the average value of the feature had a small impact on the predicted score.

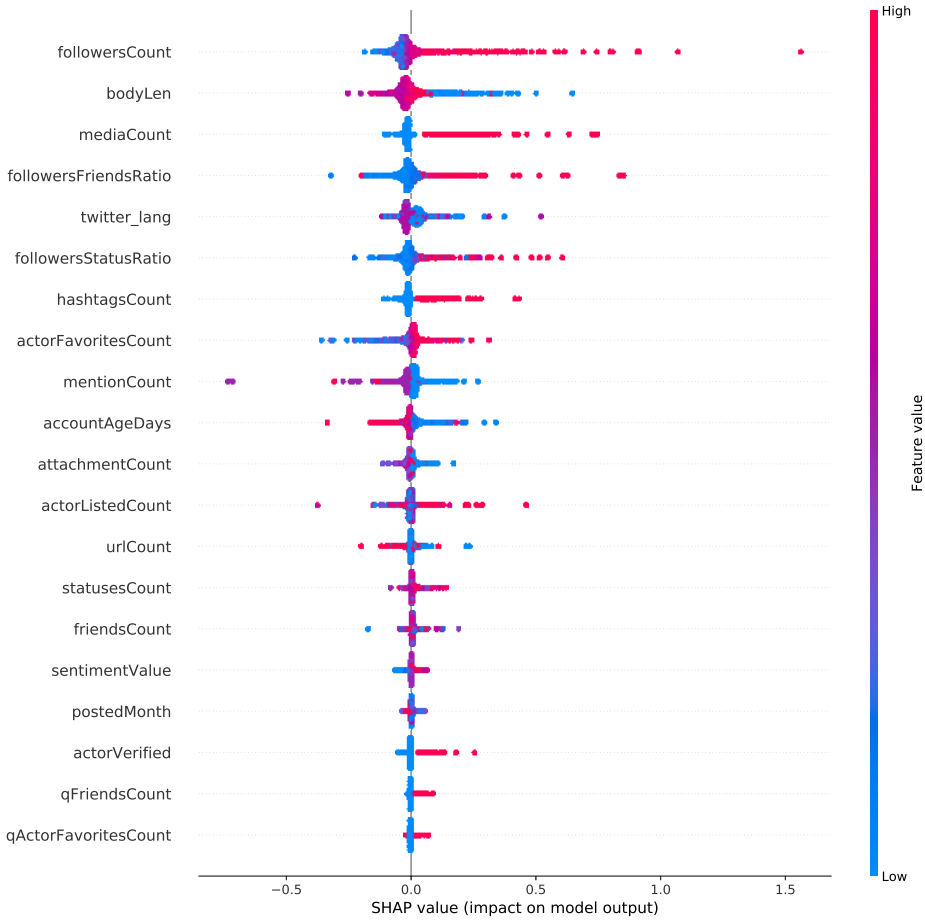


Figure 2.5: SHAP summary plot for the Virality model. High number of followers, presence of an image and at least one hashtag consistently increase the virality prediction. Among less expected findings, it is noteworthy that mentioning other users, or posting a long tweet from an old account consistently hurt the predicted virality.

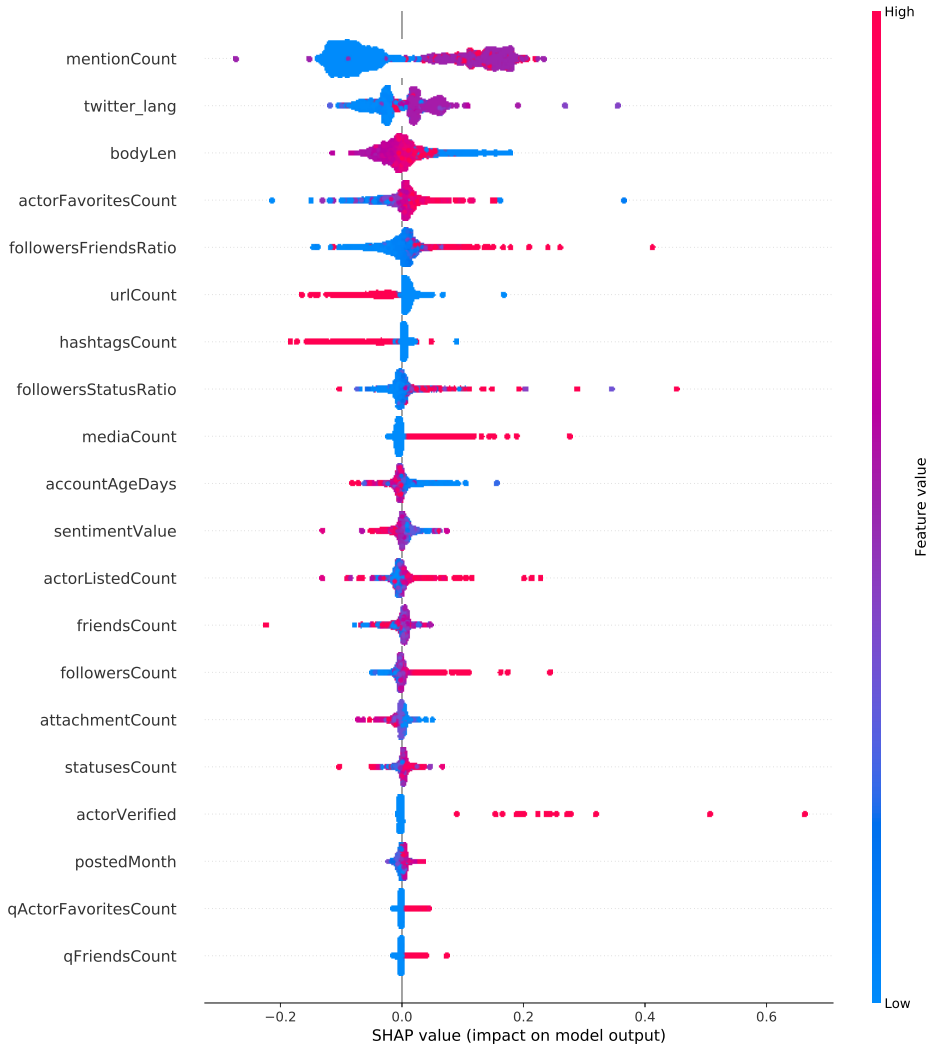


Figure 2.6: SHAP summary plot for the Response model: mentioning a limited number of accounts and attaching an image consistently increase the expected number of replies. Unexpectedly, so does the number of past tweets favorited, suggesting a reciprocity of attention on Twitter. It is also noteworthy to see the that longer body, attachment of a URL (both requiring additional effort from the responder) or hashtags (increasing visibility of the conversation, per Fig.2.5) both negatively impact the number of expected replies.

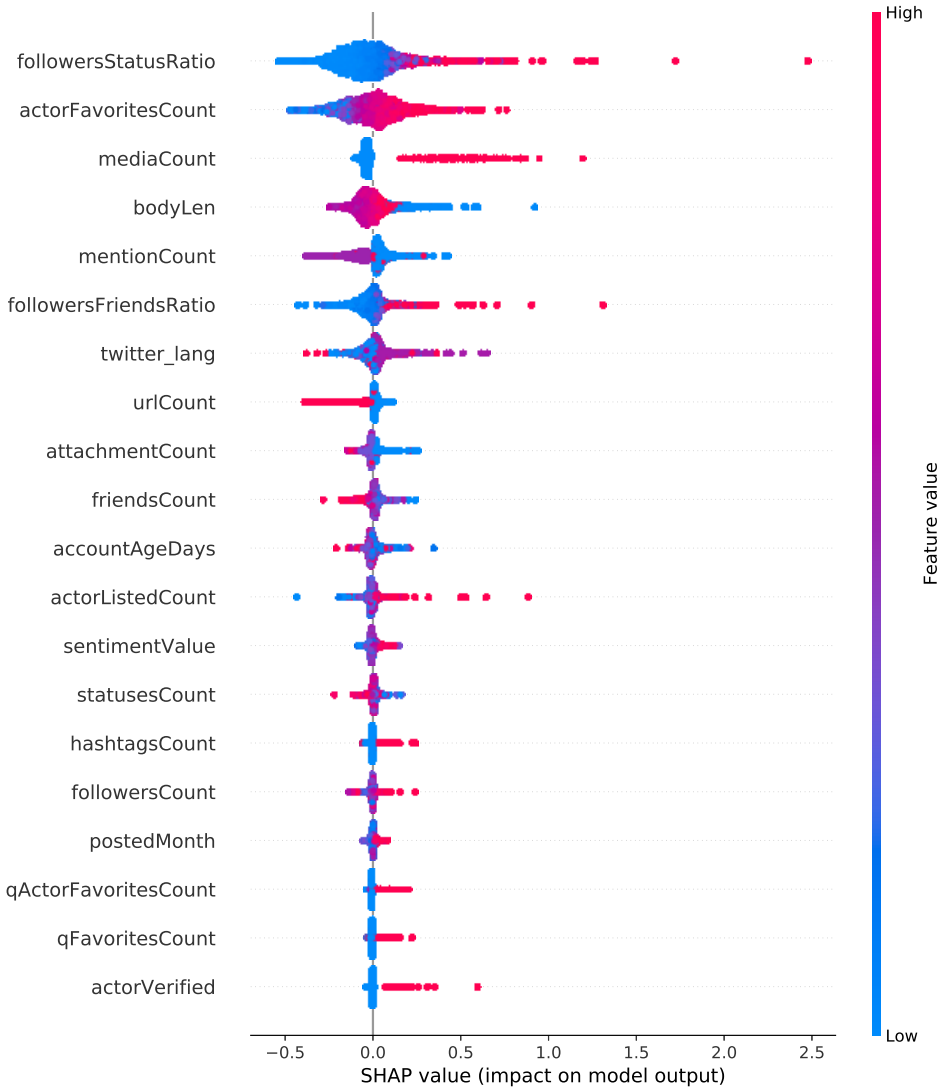


Figure 2.7: SHAP summary plot for the Popularity model: The number of followers gained with every status posted (or account followed) since joining Twitter, the number of other tweets favorited, and a presence of an image attachment all consistently rise the number of predicted likes. The importance of liking other tweets again suggests a level of attention reciprocity on Twitter. Among other findings: short body helps, mentions don't. The URL attachments again decrease the number of expected likes.

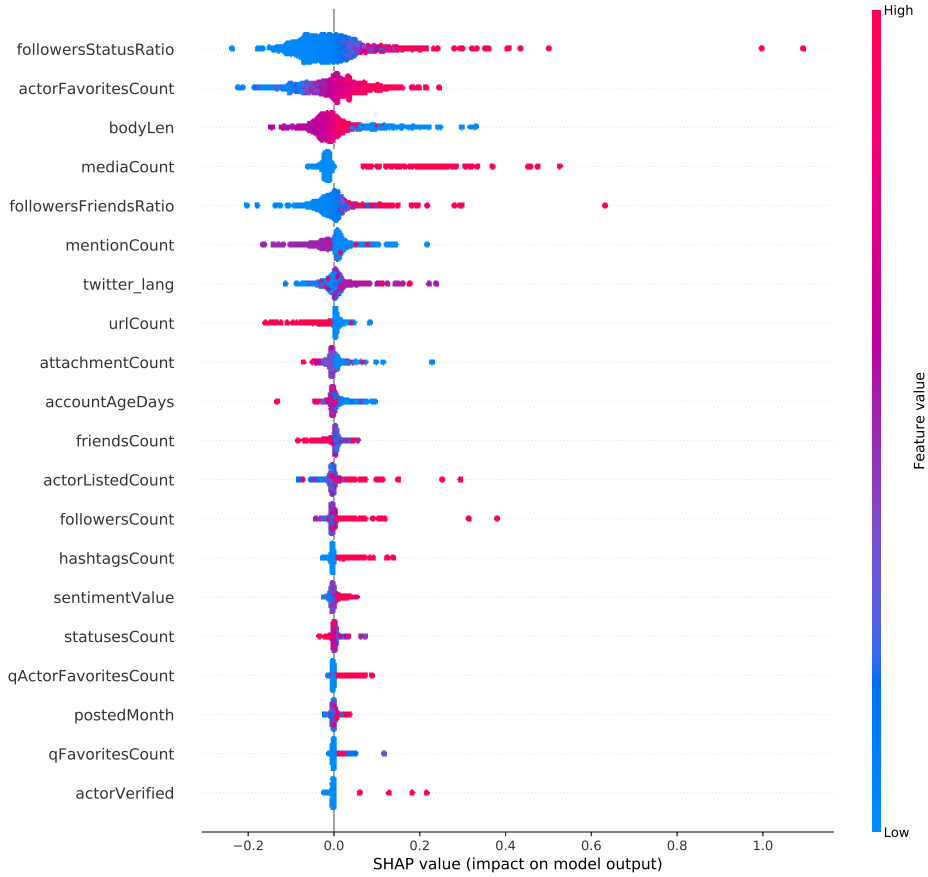


Figure 2.8: SHAP summary plot for the Engagement model. The number of followers gained with every status posted (or account followed) since joining Twitter, the number of other tweets liked, presence of an image attachment or a hashtag all increase the predicted compound engagement. Attaching a URL (pointing to arguably an additional effort, in the contrast to an instantly gratifying visual attachment) or, surprisingly, the number of accounts followed, all limit the expected compound engagement. Compound Engagement explanations conclude this project's analysis of the dynamics of attention in Twitter, yet offer an important baseline for the followup studies, in understanding topic specific (e.g., #globalwarming) audience engagement.

2.4 Beyond Twitter: Popularity Prediction on Instagram

The predictibility of social media popularity is a topic of much scientific interest and practical importance. In [RKH20] we present a new strong baseline for popularity prediction on Instagram which is both robust and efficient to compute. Instagram poses additional challenges for knowledge extraction and delivery, from data collection, analysis and operationalization standpoints. Therefore the study offers an additional opportunity to validate re-usability, beyond Twitter, of the new computational framework proposed in Chapter 3. In contrast with Twitter, the author specific attributes, proven earlier to be highly relevant for popularity prediction, are not available at the time of UGC collection, requiring dedicated collection and conflation processes before training or scoring. Moreover, attaching an image is now a prerequisite for posting, by Instagram design, thus motivating inclusion in the analysis, for the amount of potentially relevant signal carried by the visual modality.

Careful selection and extraction of visual features towards a multimodal representation of the UGC, remains as one of the more expensive open problems in the field, with five new attempts in the last October alone [CLZ+19, DMW19, DWW19, HHWY19, HLK+19]. [GO19] show significant advantages of extensive feature extraction from several sources combined with embeddings of past posts, however at the sacrifice of scalability and generality, and thus production potential. Compromises like that are still not an option within this project. To increase the potential for commercial application, I focus the feature extraction process only on signals available at the time of posting, and dedicate much of the effort to enhancing the predictive power of the visual modality with deep feature extraction. Recent advancements in computer vision offer exciting new opportunities for analyzing image attachments in a scalable way [RF18, TL19a].

In [RKH20] we expand previous work by a comprehensive ablation study of the predictive power of multiple representations of the visual modality and by detailed use of explainability tools. We explore different combinations of deep semantic features extracted from images, and their potential for transfer learning in UGC popularity prediction. We use recent pretrained models to capture various complimentary aspects of an image: concepts, scenes and objects detected, towards the richest attainable representation of a UGC at the time of posting. I then apply the GBRT approach from the Engagement study, to explore the limits of popularity prediction on Instagram, without compromising interpretability or scalability. Finally, I apply [LEL18] to calculate attribution of all the pre-trained models and predictive concepts, in the first explainable ablation study on Instagram. The results inform optimization of UGC repre-

sentation in future productive scenarios and offer new insight into the dynamics of attention on Instagram. The contributions of this study can be summarised as below:

- a scalable framework for distributed, GPU accelerated extraction of deep visual features from social images (described in Chapter 3)
- a new model for predicting popularity on Instagram, the first to achieve strong [Coh88] ranking performance while satisfying the properties of scalability and interpretability
- a comprehensive ablation study to quantify the contribution of every predictor, explained via locally accurate Shapley values [LEL18]

The report on this study is structured as follows. First, I summarize the collection of training data and outline the feature extraction. I list seventeen social features available at time of posting and introduce a new feature extraction process implemented part of this project. Then, I define the predictive task and a regression model, to be trained by the LightGBM [KMF+17] framework. The last sections focus on the results and SHAP [LL17a] explanations of a comprehensive ablation study, towards a new understanding of the dynamics of attention on Instagram.

2.4.1 Data collection

Several recent studies note that no publicly available dataset exists for Instagram [GO19, ZCZ+18, MPW18, OMW+17]. I use the data collection framework proposed in Chapter 3 to collect a new dataset. Afterwards, I follow [DMW19, RvDMW20, ZSY18, ALK16, BSG14, GO19, ZCZ+18, MPW18, OMW+17] to scrape additional features (author metadata and image attachments), which I then conflate, towards a multi-modal dataset. The resulting set consist of one million unique image posts, created in the Autumn of 2018. Relative to prior work, this volume is among the largest on both Instagram and social media platforms in general, cf. Figure 2.1. The data set is not category or user specific and as such suggests potential for generality w.r.t. all image posts on Instagram. The engagement signal collected simultaneously, included the number of likes (favorites), which I consider as the popularity signal for this study.

2.4.2 Feature extraction

Prior work proves that popularity prediction benefits from a multi-modal approach [HLK+19, DWW19, WBF18] to UGC representation. This section describes the information extracted for ever post, divided on a top level, into social and visual features.

2.4.2.1 Social features

Table 2.12 offers a summary of social features, grouped into three categories: author, content, and temporal. For each author of a post I extract the number

Author	**Skewness**	**Type**	**Origin**
followers	2.728	ordinal	original
following	3.462	ordinal	original
posts	5.183	ordinal	original
follower per post	22.368	continuous	computed
follower per following	34.635	continuous	computed
Content	**Range**	**Type**	**Origin**
filter	[0, 41]	categorical	original
users tagged	[0, 20]	ordinal	original
user has liked	[0, 1]	categorical	original
has geolocation	[0, 1]	categorical	original
language	[0, 72]	categorical	original
is English	[0, 1]	categorical	computed
hashtag count	[0, 60]	ordinal	computed
word count	[0, 519]	ordinal	computed
body length	[1, 2200]	ordinal	computed
Temporal	**Range**	**Type**	**Origin**
posted day	[1, 31]	categorical	computed
posted week day	[0, 6]	categorical	computed
posted hour	[0, 23]	ordinal	computed

Table 2.12: Summary of the social features extracted for each UGC. Based on [RKH20].

of followers, number of friends (accounts followed), and the number of posts still publicly available. In order to stabilise the variance, I log-transform these within the sample. The transformation is given as follows by first log transforming the variable:

$$y_{log} = \log(x + 1) \tag{2.18}$$

and then subtracting the mean

$$y_{trans} = y_{log} - \text{mean}(y_{log}). \tag{2.19}$$

Finally, I compute the ratios *follower per post* and *follower per following* similar to my studies on Twitter, to better capture social phenomena such as celebrity. Among the content features, I extract the filter id (Instagram offers 42 filters applicable to an image), number of users tagged in the post, whether the user has liked the post, whether there is any location data available, language id, number of hashtags, and the length of the caption measured in words and characters. I compute the *is English* flag to better separate English content from others. Among the temporal features, I first extract the time of posting: *posted date*, *posted week day*, and *posted hour*. User id and post id, are discarded immediately towards anonymity and generality.

2.4.2.2 Visual features

I deploy four pretrained neural networks within the feature extraction framework described by Chapter 3. We use the new process to capture concepts, scenes, objects, intrinsic image popularity and additional high-level features from all images collected.

- *Concept features*: We use the state-of-the-art model EfficientNet [TL19a] pre-trained on ImageNet [RDS+15] to extract concept features, informed by the 1000-dimensional feature vector of the softmax output layer.
- *Scene features*: We use Places365 ResNet-18 [ZLK+18a] to extract a variety of scene features, represented by the 365-dimensional feature vector of the softmax output layer, the 102-dimensional feature vector given by the attributes to the scenes, and a flag indicating if the scene of the image is inside or outside.
- *Object features*: We use YOLOv3 [RF18] pretrained on COCO [LMB+14] to detect 80 different objects in the image. For each object, I count the number of instances, which generates a 80-dimensional feature vector denoting the count of every object detected.
- *Intrinsic image popularity*: Here we adopt the IIPA model to directly assess the intrinsic image popularity with a single value, as proposed by [DMW19].
- *High-level features*: We extract additional high-level features represented by 2304, 512 and 2024 dimensional feature vectors from EfficientNet [TL19a], Places365 ResNet-18 [ZLK+18a], and IIPA [DMW19], respectively. All features are extracted after the last pooling layer. Since the different networks are pretrained on different data sets for different task, their internal representations of features should vary [ZLX+14].

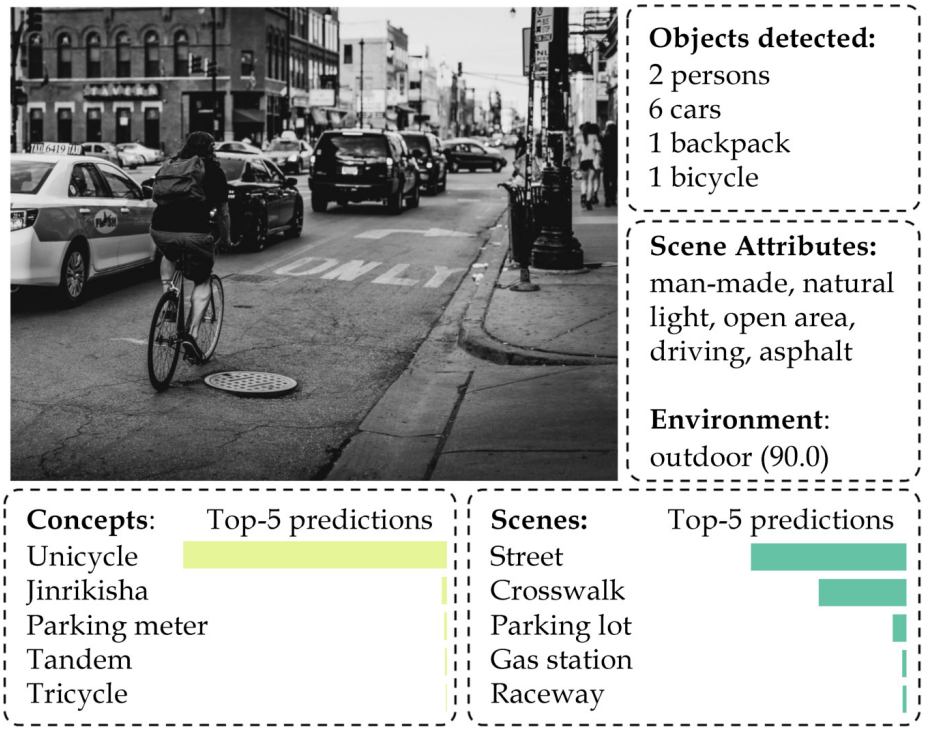


Figure 2.9: Results of the new feature extraction framework, when applied to a sample image. The associated concepts are extracted with EfficientNet, objects are detected with YOLO, the associated scenes and scene attributes including environment description (indoor/outdoor) are extracted with Places365. The image scores a natural IIPA value of 1.96 on a scale from -4 to 8. Mean IIPA score in the sample is 2. Based on [RKH20].

With the described technique, we extract a total of 1548 features representing concepts, scenes and objects, plus one value representing the intrinsic image popularity, and 4864 high-level features contributing to an expressive and comprehensive feature representation. In combination, these features advance UGC representation well beyond prior work (see Table 2.13). The extracted visual semantics are summarized by the top-10 concepts, scenes, and objects in Figure 2.10. Figure 2.9 demonstrates the extraction with results from a sample image.

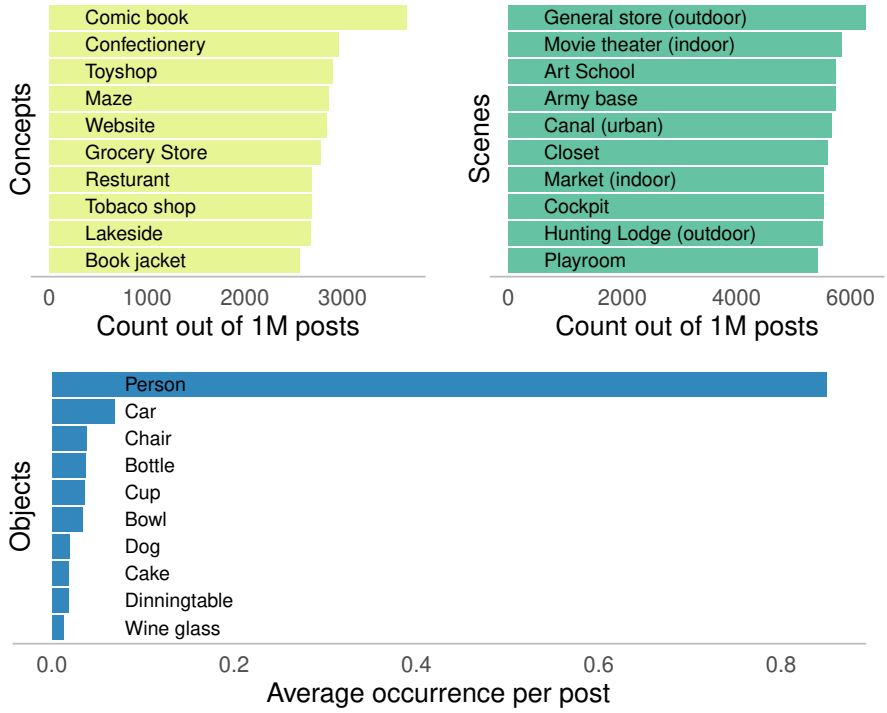


Figure 2.10: Ten most frequently predicted concepts, scenes and objects in the sample. For each category in concepts and scenes, I report the number of times it was the top prediction in the sample. Based on [RKH20].

2.4.3 Predicting popularity

In this section I motivate the basic assumptions of the popularity prediction conducted in this study, including the definition of the problem and choice of the machine learning framework.

2.4.3.1 Definition of popularity

The choice of engagement signals to define popularity in prior work varies across social networks. On Twitter focus is often on the number of retweets, but the number of likes is also used as a measurement of popularity [KL19, WBF18, KH20, ZJ19]. On Flickr and Instagram the literature is more con-

	Concepts	Scenes	Objects
[GO19]	✓		✓
[GUB+15]	✓		
[KDH14]	✓		
[MPW18]	✓		
[MRR+16]	✓		✓
[MMJ14]		✓	✓
[OFB19]	✓	✓	
[OMW+17]	✓		
[RvDMW20]			✓
This study	✓	✓	✓

Table 2.13: Prior use of concepts, scenes, and objects extracted from the visual modality. Based on [RKH20].

sistent with respect to popularity. Popularity on Flickr is measured by the number of comments [MMJ14] and clicks [CLZ+19, KLT+19], but the most common predictor is the number of views [DWW19, GUB+15, HCYH17, HLK+19, HYA16, KDH14, MMJ14, OFB19, WCZM16]. Prior work on Instagram agrees on the number of likes as the best predictor of popularity [ALK16, BSG14, DMW19, MPW18, MRR+16, OMW+17, RvDMW20, ZCZ+18, ZSY18], with few researchers using the number of comments as well [BSG14, RvDMW20]. In this study I will follow majority of the prior work to use the number of likes as the response variable.

2.4.3.2 Machine learning technique

Prior work shows many approaches to predicting popularity. [CKXM19] predict the number of mentions for a specific event; [ALK16, OFB19, WCZM16] look at the popularity over time; [OMW+17, RvDMW20] consider popularity for different brands; [MPW18] predicts popularity for different categories; [DP15, MMJ14, ZCZ+18] approach it as a binary classification problem, e.g. *popular* vs. *unpopular*; but the majority of work predict the number of likes, shares, views, etc., as a regression and ranking problem [CLZ+19, DMW19, DWW19, GO19, HHWY19, HLK+19, KLT+19, KL19, KH20, ZJ19]. Gradient boosting algorithms are used in social media popularity prediction [HLK+19, KLT+19, GO19, HHWY19, CLZ+19, KL19, KH20] due to speed, performance and explainability. Encouraged with earlier results, I continue to use the LightGBM framework [KMF+17] to address popularity prediction as a regression and ranking problem.

2.4.4 Experimental setup

In this section I summarize the technical assumptions of the experiments, including evaluation metrics and the training configuration before discussing the results of a comprehensive ablation study [RKH20].

2.4.4.1 Evaluation metrics

We focus on the most widely used UGC engagement signal on Instagram [ALK16, BSG14, DMW19, MPW18, MRR+16, OMW+17, RvDMW20, ZCZ+18, ZSY18] offered by the number of likes. We predict the log-normalised number of likes (see (2.18) and (2.19)) and choose Root Mean Square Error (RMSE) to calculate the loss during training. We follow [KL19, KH20] to compute Spearman Ranking Correlation (SRC) and R^2 as the evaluation metrics.

2.4.4.2 Training platform

In the ablation study we evaluate the predictive performance of 36 combinations of feature groups in 3-fold cross validation. It is easy to calculate that at least 108 training runs will be necessary after deep feature extraction and initial hyper-parameter tuning. The computational cost here, inherent to both feature extraction and training at scale would be prohibitive for many, thus offering a final scalability test of the analytics framework I propose in Chapter 3. The framework at this point is scaled out to three Apache Spark nodes, each equipped with a 6-core Intel Xeon CPU and NVidia Tesla V100 GPU.

2.4.4.3 Hyper-parameter tuning

I perform a very basic hyper-parameter tuning of the Gradient Boosted Regression Trees offered by [KMF+17] on the full combination of feature groups (denoted as YIEPACT) and fix these parameters across ablation experiments, to ensure fair comparison. I cap the number of leaves at 256, set the feature sampling at every iteration to 0.5 (expecting many noisy features to slow down the training otherwise), limit the number of bins when building the histograms to 255 (limit dictated by the GPU implementation [ZSH17]) and set the learning rate to 0.05.

2.4.5 The ablation study

I evaluate 36 combinations of feature groups in 3-fold cross validation, to quantify the impact of each modality on model performance, ranking power, variance captured and the overall training time. The 1M dataset is split 80/20 for training and validation. The hyper-parameters are fixed as described in 2.4.4.3. In Table 2.15 we report the SRC, RMSE and R^2 results for all the models, and illustrate further in Figure 2.11. The first 6 models are built with social feature groups only, followed by 15 models without the (promising and sourced independently) author feature group. After including the author features in training, we observe a sudden jump in performance, as the SRC increases beyond 0.5 and the RMSE decreases below 0.35. In social and behavioral sciences, ranking performance in excess of 0.5 SRC is considered strong [Coh88].

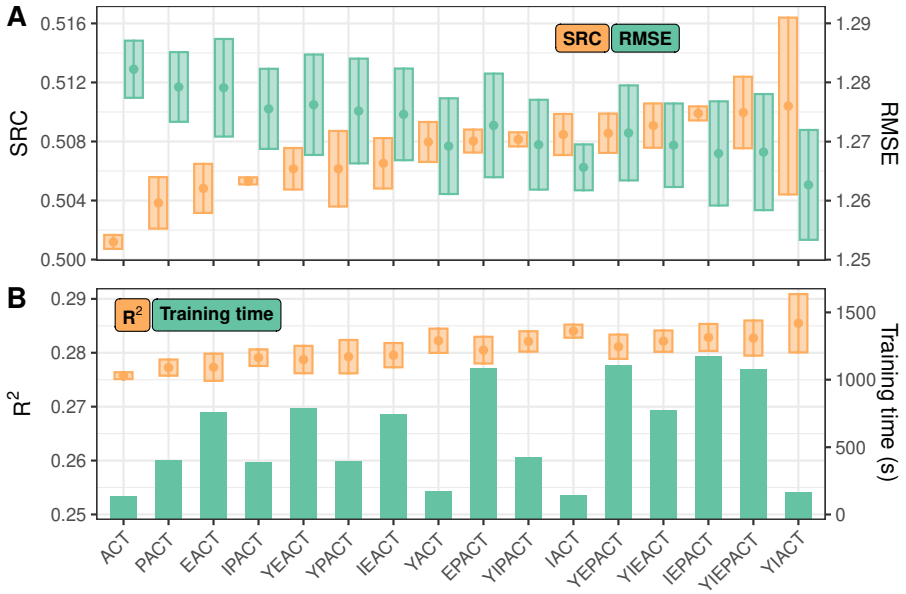


Figure 2.11: Performance for models getting a SRC higher than 0.5. The boxes show ±2 standard deviations. Model name (x-axis) explains features used: author (A), content (C), temporal (T), EfficientNet (E), Places365 (P), YOLOv3 (Y), and IIPA (I). Based on [RKH20].

The top panel of Figure 2.11 shows the performance ±2 standard deviations for 16 best models. The SRC and RMSE are inversely related as expected. The standard deviation of performance between cross-validation folds offers a conservative estimate of the standard error of the mean. YIACT has the highest SRC, but also a high standard deviation, while the model IEPACT has similar performance but is much more robust. The bottom panel in the same figure shows the variance explained R^2 and the training time. It is noteworthy that the models ACT, YACT, IACT, and YIACT were relatively fast to train, with CV time below 200 seconds. All the other models have more than four time as many features, which is reflected by the increase in training time. Considering variance explained, YIACT has the highest R^2, but IACT shows performance with much lower standard deviation.

The model IACT after a relatively short training, achieves high R^2 and a strong SRC within a small confidence interval. Therefore we propose the IACT model as a strong, robust and efficient baseline for popularity prediction on Instagram. With a longer training time (about 20 minutes) the model IEPACT becomes a robust runner-up candidate with a strong consistent SRC performance across all CV folds.

Without feature	RMSE	SRC	SRC std.	Training time
Author	1.202	0.463	0.000	1075
EfficientNet	1.158	0.509	0.000	421
Places365	1.158	0.509	0.001	772
YOLOv3	1.157	0.510	0.000	1170
IIPA	1.159	0.509	0.001	1105

Table 2.14: Ablation study by feature group removal. That author features are more important than the visual features. The removal in EfficientNet gives the largest decrease in training time with almost no reduction in performance. Based on [RKH20].

	SRC		RMSE		R^2		Time
Features	μ	σ	μ	σ	μ	σ	(sec)
T	0,261	0,001	1,306	0,001	0,086	0,001	30
C	0,305	0,002	1,291	0,001	0,108	0,001	50
A	0,349	0,002	1,266	0,001	0,141	0,001	124
CT	0,417	0,001	1,231	0,001	0,188	0,000	71
AT	0,425	0,001	1,219	0,002	0,204	0,001	134
AC	0,426	0,000	1,216	0,001	0,207	0,000	151
CT							
YCT	0,433	0,000	1,222	0,001	0,200	0,000	101
ICT	0,435	0,001	1,219	0,001	0,204	0,000	61
YICT	0,444	0,001	1,214	0,001	0,211	0,001	101
PCT	0,452	0,001	1,210	0,001	0,216	0,001	362
ECT	0,455	0,000	1,208	0,001	0,219	0,001	719
YPCT	0,456	0,000	1,207	0,002	0,220	0,001	388
IPCT	0,456	0,000	1,206	0,001	0,221	0,001	356
YECT	0,457	0,000	1,206	0,002	0,221	0,001	740
IECT	0,458	0,001	1,205	0,001	0,222	0,000	737
YIPCT	0,459	0,000	1,204	0,001	0,224	0,001	375
EPCT	0,460	0,001	1,205	0,001	0,223	0,000	1061
YIECT	0,461	0,000	1,204	0,001	0,224	0,001	742
YEPCT	0,461	0,000	1,204	0,002	0,224	0,001	1030
IEPCT	0,462	0,001	1,202	0,001	0,226	0,001	1087
YIEPCT	0,463	0,000	1,202	0,001	0,227	0,001	1075
ACT							
ACT	0,501	0,000	1,163	0,001	0,276	0,000	136
PACT	0,504	0,001	1,162	0,001	0,277	0,001	398
EACT	0,505	0,001	1,162	0,002	0,277	0,001	754
IPACT	0,505	0,000	1,160	0,001	0,279	0,001	389
YEACT	0,506	0,001	1,160	0,002	0,279	0,001	785
YPACT	0,506	0,001	1,160	0,002	0,279	0,002	394
IEACT	0,507	0,001	1,160	0,002	0,280	0,001	741
YACT	0,508	0,001	1,158	0,002	0,282	0,001	172
EPACT	0,508	0,000	1,159	0,002	0,280	0,001	1081
YIPACT	0,508	0,000	1,158	0,002	0,282	0,001	421
YEPACT	0,509	0,001	1,159	0,002	0,281	0,001	1105
YIEACT	0,509	0,001	1,158	0,001	0,282	0,001	772
IEPACT	0,510	0,000	1,157	0,002	0,283	0,001	1170
YIEPACT	0,510	0,001	1,157	0,002	0,283	0,002	1074
YIACT	0,510	0,003	1,155	0,002	0,285	0,003	160

Table 2.15: Quantitative evaluation of each feature group's contribution to model training and predictive performance.
Abbreviations: author (A), content (C), temporal (T), EfficientNet (E), Places365 (P), YOLOv3 (Y), and IIPA (I). Based on [RKH20].

2.4.6 Explaining popularity, discussion

The choice of an explainable GBM framework [KMF+17] for training of every model in the ablation study, creates an excellent opportunity to compare the explanations. I apply Shapley Analysis from the *shap* library [LEL18], still considered SOTA in explaining GBMs. Due to high computational cost, I use the analysis framework proposed in Chapter 3 to parallelize and distribute the workload. Single Shapley value quantifies the effect on prediction, which is attributed to a feature. Two properties of these values make them ideal for explaining our ablation study:

- **Consistency and local accuracy:** Even if I change the model so that a feature has a greater impact, the attribution assigned to that feature will never decrease. Features missing in the original input (i.e. removed in ablation) are attributed no importance. The values can be used to explain single predictions as well as to summarize the model.
- **Additivity of explanations:** Summing the effects of all feature attributions approximate the output of the original model. Additivity therefore enables aggregating explanations, e.g., on a group level, towards an accurate and consistent attribution for each of the modalities in the study.

In this section I aim to advance the understanding of the dynamics of attention on Instagram, via SHAP explanations of the popularity ablation. I structure the discussion by three phases of the analysis across models. In the first one we compare the overall impact of feature groups and how it changes with addition of other features. In the second one we identify the top 30 (out of 1548) individual features with the highest impact on predictions. In the third one we compare the average positive and negative impact of top social features. The last analysis illustrates how the attributions of top visual features change with an addition of other feature groups. To the best of my knowledge this study represents a new level of scrutiny, in explaining online popularity on Instagram.

2.4.6.1 Feature groups

The Figure 2.12 shows an average absolute SHAP value for each feature group for each model along with the corresponding SRC performance. The base model CT trained on *Content* and *Temporal* features achieving SRC of 0.417 is displayed in the upper left corner. The content features affect the prediction more than the temporal ones, since the content bar is higher than the temporal bar.

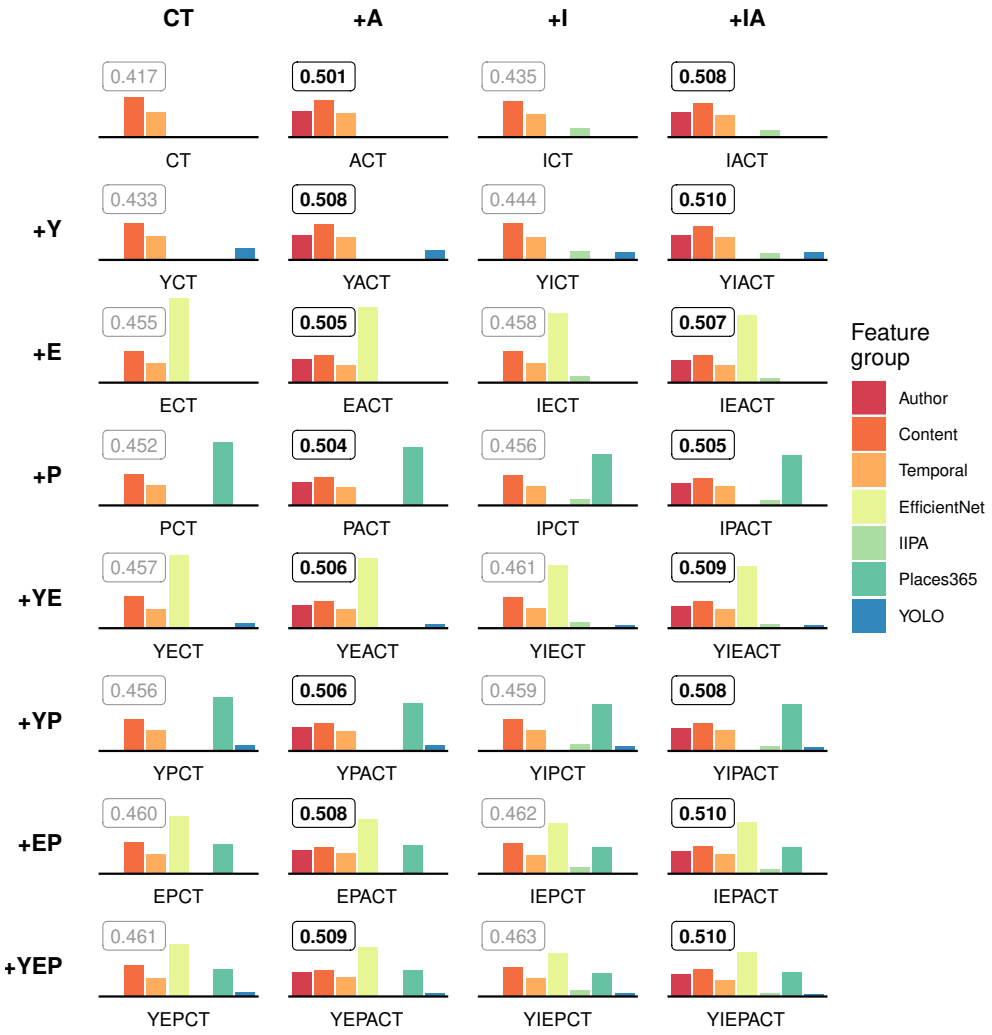


Figure 2.12: Average absolute SHAP value aggregated within each feature group used by the models. Upper left bar plot shows the base model CT consisting of *Content* and *Temporal* features. For the three columns, *Author* (A) and *IIPA* (I) features are added, and for each row the groups *EfficientNet* (E), *Places365* (P), and *YOLO* (Y) - corresponding to concepts, scenes, and objects respectively - are added. For each model the Spearman's Rank Correlation is shown in the box. Based on [RKH20].

Author features are essential: The columns to the right in Figure 2.12 include author features (A), IIPA (I) and a combination of the two (IA). When we examine the first row with the base model CT in common, we observe that adding IIPA features results in increase of the model's performance from 0.417 to 0.435 SRC. It is noteworthy that adding A offers much higher boost to performance, with SRC at 0.501. In fact, considering all the models in the second and fourth column, we observe that the author features are indeed necessary to obtain strong [Coh88] ranking performance (SRC above 0.5).

EfficientNet has the largest cumulative impact: In the rows below the base model CT in Figure 2.12, the different semantic concepts, scenes, and objects are added to the model from *EfficientNet* (E), *Places365* (P), and *YOLO* (Y) respectively. When comparing the three models YCT, ECT and PCT, we observe that E exhibits the largest effect on the predictions. In the lower half of the figure, we show the models combining more visual features, yet E maintains the largest effect. Indeed EfficientNet features have the largest effect on the predictions of all the models examined. However, we should note that E has 1000 features, whereas P and Y only have 486 and 80 resp. Despite E showing the largest cumulative impact, a single feature from P and Y might contribute more than an average feature from E. We hypothesize that many of E features cancel each other out and revisit this hypothesis in Figure 2.15.

Visual semantics are correlated: When adding combinations of visual semantic groups, we observe a decrease in attribution for a single group, e.g. in YEPCT the attribution of E, P and Y is lower than in other models in this column (ref. Figure 2.12). At the same time, we note the SRC increases every time new features are added to the model, indicating that the different feature groups are complementary. However, the decrease in different attributions coinciding with the increase in SRC also suggests that the groups are somewhat correlated and that the model might learn a better representation where some of the features within the groups are disregarded. This illustrates a synergy between the groups, with how features can be substituted after including others. Above observations can be validated across the columns. When we examine the three other columns, we observe in the second row that the effect of Y decreases after adding A, I, and IA. The same is true for E and P as the model is combining the visual semantics. In fact, the more features we combine the lower is the contribution from each feature group. In particular, the largest model YIEPACT shows the lowest average attribution for each feature group.

Author features increase attribution of detected concepts: The second column of Figure 2.12 illustrates the impact of adding the author features (A) to the base model CT. We observe a sudden increase in the performance reaching a SRC at 0.501, which is 0.04 higher than the best model YEPCT from the first column. We have already observed performance increase obtained by adding author features, so the following examines the effect of A on the visual semantics. It is noteworthy that models with EfficientNet features (E) always give the same or better performance than Places365 features (P) across all models, e.g. YIEACT has a higher SRC than YIPACT. If we examine the models without A starting with the first column, we see that the increase in performance is higher when adding E or P instead of Y, e.g. the model EPCT achieve a higher SRC than both YECT and YPCT. Same patterns are seen in the third column. However, if we examine the models with A starting with the second column the pattern is more cluttered, since YACT achieves a higher SRC than both EACT and PACT. Moreover, we see that adding either E or P to YACT results in a decrease in performance, but adding all them in YEPACT gives the highest performance in this column. Furthermore, we do see that the combination of EP in EPACT achieves the same performance as YACT.

Towards optimal representation: From the observations we hypothesise that when adding a single semantic group to ACT, YOLO is preferred for the highest performance, but if adding two is possible then E and P will be preferable. There seems however no advantage in predictive performance to using EP instead of Y. Finally, even though both YEACT and YPACT have lower performance than YACT, adding all three visual semantics to YEPACT offers a marginal benefit. The fourth column validates these hypothesis, where again Y as a single feature is better than E and P, but adding the combination EP gives similar performance to adding Y. However, here no significant performance gain is observed by combing YIACT and IEPACT into YIEPACT. All these three models achieve the highest observed SRC at 0.510. The cumulative attribution highlights the predictive power of objects together with the authors features, but also shows how the combination of concepts and scenes remains powerful without author features.

2.4.6.2 The Top-30 features

In this section I investigate the features with the highest overall attribution in the ablation study. I discuss the top-30 most prominent features based on the average absolute SHAP value across all models. More precisely, I sum the average absolute SHAP value for each feature across all models and then divide by the number of times the feature was used by the model.

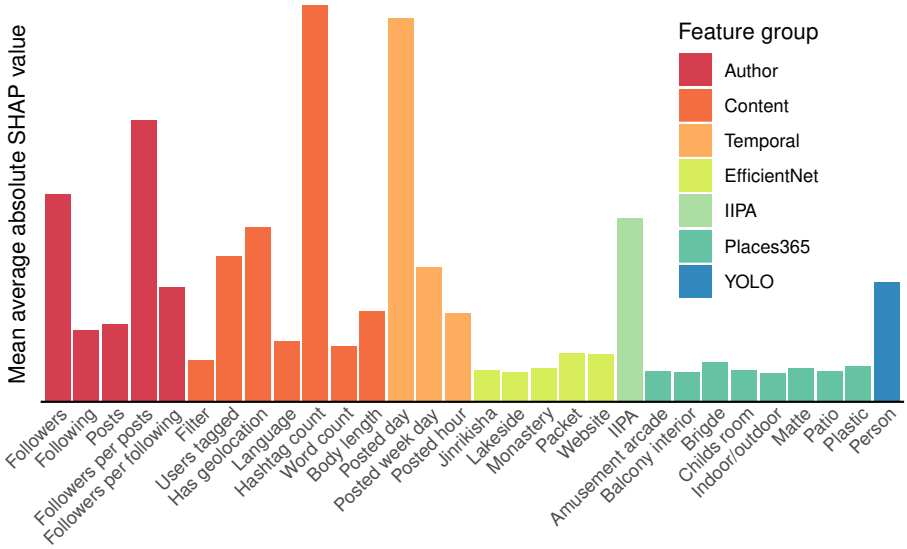


Figure 2.13: Based on [RKH20]. Average absolute SHAP value for top 30 features. The features are ranked by the highest average absolute SHAP values across all models, then normalized by the number of times the feature is used, for fair comparison. It is noteworthy, that all 16 social features are among the top ones.

In Figure 2.13 the top-30 features are shown coloured after each feature group. The two features *hashtag count* and *posted day* have by far the largest average absolute SHAP value and as such affect a prediction the most. The author features *followers* and *followers per post* come right after with high attribution as well. The two computed ratios *followers per post* and *followers per following* introduced in 2.4.2.1 prove more impactful than the two original features *following* and *posts*. The three temporal features all have a high effect on the prediction, illustrating that the time of posting on Instagram has significant impact on expected popularity. This is a striking difference with popularity on Twitter explained by Figure 2.7. The content features *users tagged*, *has geolocation* show also a relatively high effect. Among the visual features, IIPA and *Person* have the largest effect, in fact comparable with the social features. Otherwise most of the visual features have a smaller effect. In the following, I will therefore investigate the effect of the social and visual features separately.

2.4.6.3 The social features

The social features are explained individually using SHAP values. The positive and negative means for the social features are visualised in Figure 2.14. I aggregate the SHAP values into the mean of all positive and all negative SHAP values separately. In that way, I both preserve the sign and deviation of the attributions. Otherwise, the values of the opposite sign would cancel each other out in a regular mean calculation across predictions.

Hashtag count and posted day are important: In Figure 2.14 the base model CT trained on content and temporal features indicate that *hashtag count* and *posted day* are good discriminators. There are two explanations to note: they have high positive and negative SHAP means, and the magnitude of the positive and negative mean is similar, meaning these features impact a prediction in both positive and negative direction. The number of *users tagged* also has a high impact on the prediction, and the effect is mostly in a positive direction since the positive mean is of larger magnitude than the negative mean. Consequently, the discriminative power is lowered than that of the two aforementioned. Next, *has geolocation* appears to be a good discriminator, *filter* mainly affect the prediction in a negative direction, whereas *language* mainly affect the prediction in a positive direction. Overall attribution across the social features in this figure shows similar trends to Figure 2.13, discussed in the previous section.

Relationship between language and visual semantics: If we consider the first column in Figure 2.14, only small changes are observed down the rows. The size of the bars is decreasing slightly as we add visual features, e.g. *word count* is larger in CT than YEPCT. Adding objects (Y) only seems to have minimal effect on the bars without changing the relative distribution, whereas adding concepts (E) and scenes (P) gives an increase in the positive mean of *language*. In fact, all the features have lower impact in YEPCT than in CT except *language*, which contributes more. This shows that *language* is more important when visual semantics are added to the model. The other columns validate the observation and lead us to hypothesize, that the visual predictors of popularity vary significantly across cultures (arguably captured to an extent by languages).

The impact of caption decreases with visual features: Considering Figure 2.15, if we compare the models in the first row with the models in the last row, attribution of the feature *word count* is reduced after including caption. Word count is the number of words in the caption, as a feature it becomes less important when visual features are included in prediction. Such connection between the visual features and basic textual features like the number of words, suggest that the visual information can partly substitute predictive signals in text.

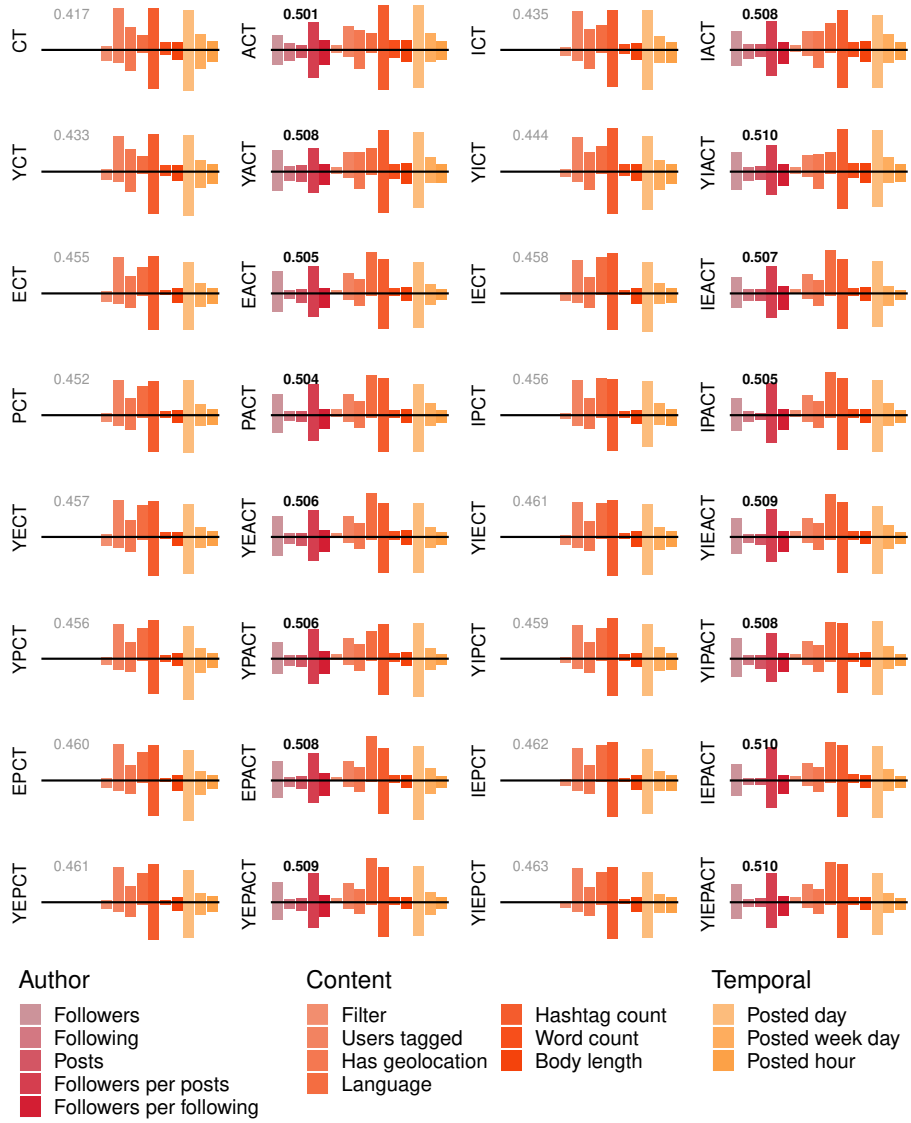


Figure 2.14: Average positive and negative SHAP values for the 15 most important social features show significant differences between the 32 models in the ablation study. Based on [RKH20].

The high impact of author features: The second column of Figure 2.14, describes the models with author features added. Comparing CT with ACT models, a large decrease in the attribution of *users tagged* is observed indicating that this feature is less important when author features are included. The same observation can be made in the other rows of these two columns. The features *followers* and the two computed ratios *followers per post* and *followers per following* have a relatively large effect on the prediction. It should also be noted how these two features become more important as visual semantics are added. The difference in attribution of *followers* and *follower per post* between ACT and YEPACT models is substantial.

The impact of language depends on IIPA: In the third and fourth column of Figure 2.14 IIPA is added to CT and ACT respectively. Like E and P, IIPA also increases the positive mean of *language*, as illustrated by the difference in CT vs ICT. This is also seen for other rows though the increase is smaller, due to the increase from E and P. Therefore, we observe that language is more important along with IIPA, suggesting that the very definition of popularity differs across cultures.

Visual features have a small impact on social features: Overall, only small changes are observed across the models in Figure 2.14, indicating that the visual features have only a small effect on the impact of the social features on a prediction. If we compare the models in the first row with the models in the last row, the features *language* has increased and *word count* has decreased. If we compare ACT with YIEPACT, we can observe that the majority of the features have a smaller impact. The SHAP attribution of *word count* is reduced but those of the two author features *followers* and *followers per post* are unchanged, and the attribution of the content feature *language* is actually larger. This suggests that author features are important regardless of the visual information, that *language* might capture some sort of user segment, and that *word count* and visual information are highly related.

2.4.6.4 The visual features

The average attribution for the visual features across ablation models is illustrated by Figure 2.14. Again I aggregate the SHAP values into the mean of all positive and all negative SHAP values separately. The first rows are sparse since these models primarily consist of social features, e.g. CT is trained only on content and temporal features. It should also be noted that the scale for the visual features is different from the social features therefore Figure 2.13 with top-30 features should be examined for comparison of social and visual features. This section explores the attributions of visual features alone.

IIPA and people features are important: Figure 2.15 clearly identifies *IIPA* and *person* as the two most impactful visual features, with their attribution illustrated by the largest bars. It is noteworthy that the two features become less important when other features are added to the model. By comparing YCT with YACT, it is seen that the added author features seem to decrease the importance of *person*. Similar reductions are observed when the visual features I, E and P are added - in fact, the smallest effect from *person* is observed in the largest model with all visual features, i.e., YIEPACT. When examining *IIPA*, the same trends are observed: both the addition of author features and the other visual features reduce the effect of IIPA. Again, *IIPA* has the lowest effect in the largest model YIEPACT.

Bridge is an important scene. When considering the features of P in PCT and PACT as illustrated by Figure 2.15, we observe how the addition of author features reduces all the features of P. IIPA affects the features of P in the same way. If we add both author features and IIPA at the same time (IPACT), a relatively large decrease in the two features *amusement arcade* and *balcony interior* is observed. Adding E and Y also results in large reduction in attribution of the P features, and it is actually observed that the three features *amusement arcade*, *balcony interior* and *childs room* have almost no effect on a prediction in the large model YIEPACT. By comparing PCT and YIEPACT, it becomes apparent that only the feature *bridge* is of a similar importance compared to the small model PCT. We propose that *bridge* is a scene predictive of popularity.

There is a connection between concepts and places: When examining Figure 2.15 We observe a connection between the concepts (E) and the scenes (P). When we compare the bars of E in ECT with E in the model EPCT, where the scene features are added, we notice that the concept bar *lakeside* has disappeared. Since *lakeside* indeed is a place, we hypothesise that the features from Places365 substitute for *lakeside*. Finally, we also observe how the features from both E and P are shrinking, when the two feature groups are combined. The observation is validated across the columns.

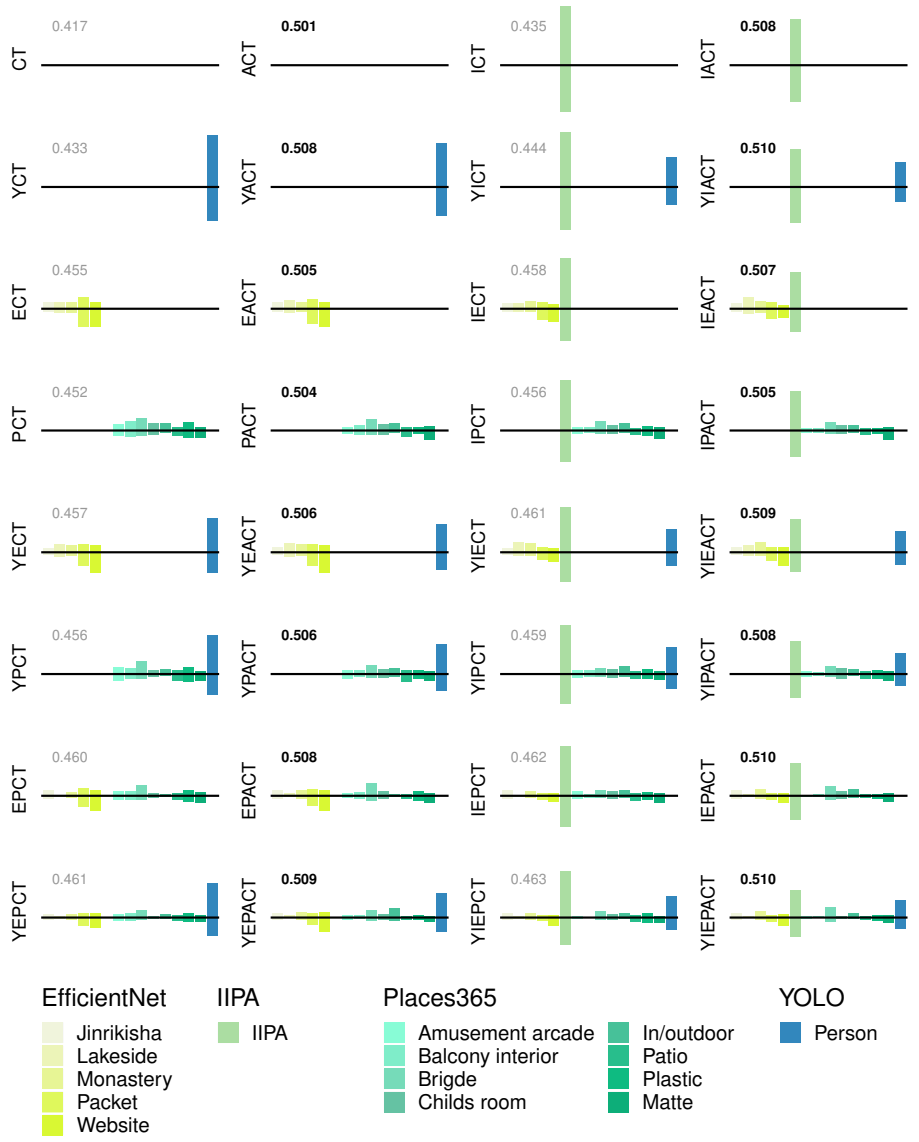


Figure 2.15: Average positive and negative SHAP values for the 15 most important visual features show subtle and significant differences between the 32 models in the ablation study.

2.5 Conclusions

In this chapter, I have summarized my results in studying the dynamics of attention at scale. Online social networks like Twitter and Instagram today aggregate traces of our collective attention at an unprecedented scale. I design, implement and then leverage a new big data framework (proposed in the following chapter) to scale up to the challenge. I use the framework to position the studies in this chapter among the largest studies on social media to date (see Figure 2.1). At this scale, I prove it is possible to rival state-of-the-art results without compromising on explainability, robustness or privacy compliance [KL19].

Non-linear advanced ML algorithms like deep neural networks and gradient boosted machines are among the most successful methods used to date. GBRT specifically offer an excellent opportunity for strong explanations (i.e., accurate explanations of a strong model). The models of virality, engagement and popularity I deliver throughout my project for both Twitter and Instagram, are the first to achieve strong [Coh88] ranking performance in a robust and explainable way. My GBRT approach combining features available early with sentiment score and high accuracy ground-truth achieves state-of-the-art results on multiple benchmark datasets. The compound engagement model, in particular, is the first to explain half of the variance with features available early, and to offer strong [Coh88] ranking performance simultaneously.

In [KH20], we examine and consolidate a diverse set of content engagement metrics from Twitter. The correlations discovered allow us to propose a new, more holistic, one-dimensional engagement signal. We then show it is more predictable than any previously investigated influence predictor. I propose the ability to engage the audience as a new, more holistic baseline for social influence analysis. We share the compound engagement workflow and parameters (Eq. (2.13) and Table (2.7)) to ensure reproducibility and inspire future work on engagement modeling. I hypothesize that this alternative metric could alleviate the negative impact of diffusion-based influence maximization, on our collective attention, well-being and by extension the democratic process itself.

In [RKH20], we address the hard problem of multi-modal popularity prediction using population models on Instagram. We employ deep neural networks to pursue a rich UGC representation, prerequisite to advancing our understanding of popularity. We conduct a comprehensive ablation study, including transfer learning to represent visual semantics with the explainable features concepts, scenes, and objects. Through ablation, we inform feature fusion towards an optimal UGC representation, validated by a strong and robust ranking performance. The study delivers the first strong popularity ranker to satisfy the properties of robustness and interpretability simultaneously.

2.5.1 Future work

The models are ready for production with immediate application to social media monitoring, campaign engagement forecasting, influence prediction, maximization, and curating user feeds. The analysis suggests immediate avenues for further inquiry. The virality study suggests that the diversity of information actively consumed by an individual might be predictive of social influence. The Instagram study suggests that visual predictors of popularity can vary across cultures. The engagement study only begins to interpret the Shapley explanations computed for each model. Further study of interactions of SHAP attributions with the language feature alone, can offer new insight into culture-specific dynamics of attention. The summary explanations, however, already offer a necessary baseline, if one was to explain topic-specific (e.g., #globalwarming) engagement at scale. These are only some of the hypothesis to exceed the three-year duration of my Industrial Ph.D. project.

Chapter 3

Big Data Engineering

big data "extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions" (Mashey, 1990)

the world "the world is one big data problem" (McAfee, 2012)

3.1 Introduction

Extracting knowledge from big data, due to its high volume, velocity, and variety (referred to as the 3V's since [Lan01]) is hard. Delivering insights from social big data, to customers globally, is even harder. The personal nature of the records makes social big data subjective to many privacy concerns and laws. The European General Data Protection Regulation (GDPR and ISO/IEC 27001), which came in force during this project, makes social network analysis more difficult to use in business, where dependant decisions need to be retraceable (explainable) on-demand [HBPK17]. Privacy and transparency remain among the biggest challenges in extracting knowledge from social networks [BOJC16, SP18]. Ensuring reliable (accurate) and explainable analysis, while respecting user privacy, remain conflicting goals and open research issues individually. Consequently, substantial effort in this industrial Ph.D. project has

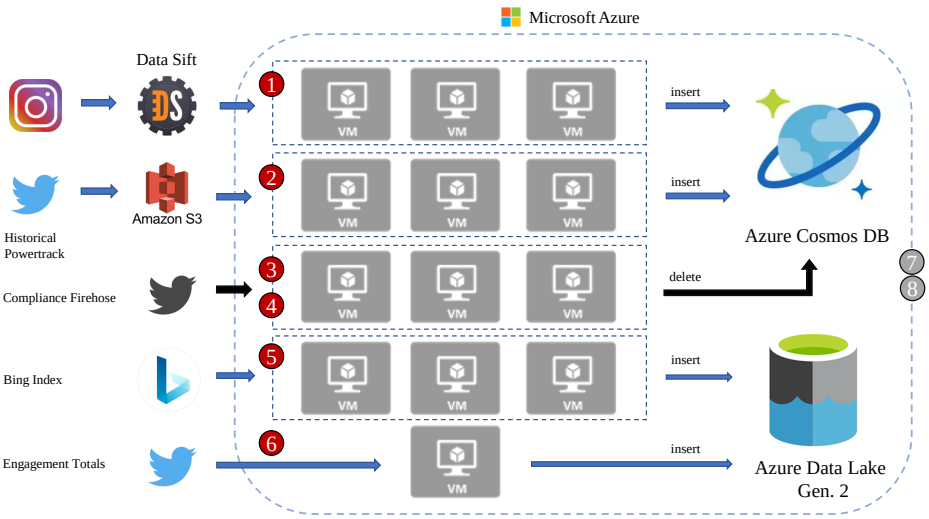


Figure 3.1: Data collection and storage topology

been directed to big data processing, privacy compliance, explainability, and operationalization. This chapter summarizes key architectural choices I have implemented part of this project, to deliver in 2019, state-of-the-art engagement analysis, to over 8000 private and enterprise customers around the world.

3.2 Data collection overview

None of the data analyzed in Chapter 2 has been available a the time of starting this project. A diverse set of APIs, however, was gradually made available for consumption, part of Microsoft agreements with Data Sift, Twitter, and subsequent collaboration with the Bing Predicts team. Figure 3.1 offers a topology of data collection pipelines implemented part of this project, to leverage these APIs. Each of them performs a unique role in maximizing signal-to-noise ratio for social engagement modelling tasks, under scalability and privacy constraints. The motivation to implement the data collection pipelines outlined in figure 3.1 can be summarized as follows:

- Retrieval and enrichment of raw User Generated Content (UGC) from Instagram and Twitter

- Application of all privacy requests issued for the collected UGC, since the first data request until the end of this project
- Retrieval and collection of engagement signals available for the collected UGC, for machine learning supervision
- Consolidation and indexing of the collected UGC for high-volume privacy requests and feature-extraction requests issued simultaneously

The following sections offer a detailed description of the individual pipelines, with a focus on the design choices contributing directly to achieving the above goals. The overall intention of the chapter remains to document the painstaking dependency of social data science on big data engineering, but also to highlight the many opportunities to advancing common data science tasks with the latest technology available.

3.3 Twitter analysis in times of GDPR

The introduction of strict privacy laws and regulations during this project motivates substantial additional workload, unaccounted for at the time of project proposal in 2016, yet prerequisite for any modelling tasks. Privacy compliant analysis in this project is defined as one that considers publicly available data only. Any document no longer public as a result of a privacy request, needs to be removed from the storage within a legally approved grace period of 30 days. The responsibility of applying every incoming privacy request to datasets controlled by Microsoft is delivered by the Compliance Pipelines (3) and (4). The design and implementation of these pipelines is described in the following sections. The purpose of this section is to offer the regulatory context, for strict order and time constraints on the data collection tasks, summarized by figure 3.2:

- **The grace period** is defined as the maximum delay in the application of a relevant privacy request, or the number of days since a particular dataset has been exposed to all privacy requests applicable. This project is granted 30 days.
- **Data transit** The volume and velocity of incoming privacy requests demand the target datasets to be exposed at rest, after indexing at the destination. Every hour spent on ETL, i.e., Extraction of the data from Twitter History, Transformation, and Loading (indexing at the destination) is considered transit, reducing the remaining grace period.

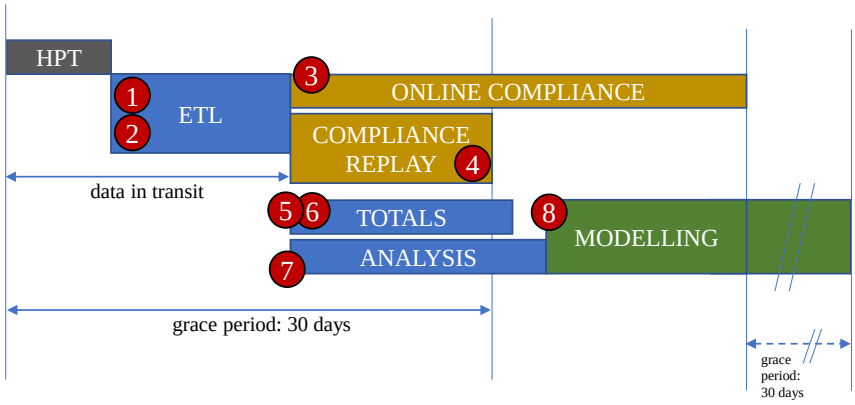


Figure 3.2: Actual time available for modelling relative to data engineering and privacy compliance workload

- **Compliance Replay** The number of privacy requests issued by users globally, and potentially applicable to data in transit, grows by an average of 2000 per second. Therefore one week of transit operations creates a compliance debt of at least 1.2 billion requests. The responsibility of the Compliance Replay (4) pipeline is to retrieve and process these requests within the grace period. Failure to do so would require deleting the data within 30 days from HPT request, or after 15 days left for analysis since transit.

- **Online Compliance** Data already indexed at a destination can and should be exposed to live privacy requests simultaneously with historical. The common Online Compliance Pipeline (3) shared across projects at Microsoft enables modelling beyond the grace period.

- **Simultaneous feature extraction** Extraction of anonymous metadata by pipelines (5, 6, 7) for documents already at rest, is executed in parallel with Compliance Replay, to ensure the modelling phase (8) can start as soon as possible.

- **Modelling** of social engagement, can only begin after all relevant data is indexed at the destinations, which includes supervisory signal delivered by pipeline (5, 6) and advanced content representation from the pipeline (7). The time for modelling ends with the last day of the grace period, after which all UGC is deleted.

The asymmetry between time spent on engineering and data science tasks affects big data researchers around the world. In 2015, [HS15a] offered an elaborate explanation of this asymmetry, through the lens of hidden technical debt. Since the time available for data science tasks is legally dependant on the scalability of data engineering foundations, these investments receive top priority in the first half of the project and a detailed description in this chapter. To consider the analysis reported in Chapter 2 as a result of three years of work would be false, for a substantial effort in this project was required upfront to pay the technical debt of compliant engagement modelling at scale.

3.4 Data storage and serving

The volume, velocity, variety, fragmentation, and sensitivity of the data analyzed in this project impose strict requirements on data warehousing in particular. Over the last four decades, relational databases such as MS SQL Server, have dominated data warehousing solutions, for many good reasons, including space efficiency or preventing data duplication, via normalization. Today they are also understood best, which could explain a wide adoption in many commercial and academic projects. However, relational databases also have significant shortcomings. The enforcement of schema and selective indexing requires from an administrator to flatten all data before storage and foresee all analytic workloads ahead to prevent expensive schema changes. The cost of unforeseen feature transformations (e.g., dependant on non-indexed fields) will directly impact the number of experiments ran, or the hypothesis space explored. The necessity to perform schema update in a production environment with the ingress of 50M documents per day (10% of Twitter's average daily rate), can be catastrophic. [BGU16] emphasizes scalability as the other major shortcoming of traditional SQL databases. Unless at the time of provisioning of the server, the administrator can foresee and accommodate the total data volume, the incoming big data, by 3V definition, is guaranteed to cause an expensive reconfiguration and delay. The commercial goal of this industrial Ph.D. project is to produce general and scalable models to power online services on a global scale. In order for the model to be relevant globally, it has to generalize across cultures and languages. Intuitively the study to deliver such models is one learning from behavioral data representative of the global population. The central hypothesis of this project is that strong general models are achievable, based on a large-scale sampling of behavioral data from online social networks. These assumptions dictate data warehouse solution capable of massive storage and scalability from the beginning.

3.4.1 Content Store: Azure Cosmos DB

Azure Cosmos DB is a proprietary globally-distributed NoSQL database "for managing data at the planet-scale" launched by Microsoft in May 2017. After nearly three years of rapid growth in market share, it is also well documented [Gua18]. Below summarizes the key strengths of Cosmos DB, relative to traditional DB management systems, in the context of this project:

- **Not Only SQL**: CosmosDB does not enforce a schema, nor flattening of incoming data. Nested JSON objects representing UGC activities, can be stored without transformation dedicated in NoSQL collections.
- **Turn-key scalability**: Cosmos collections support dynamic re-partitioning (scaling). Each partition is backed by a dedicated SSD or NVMe drive. Re-partitioning is triggered automatically when incoming data exceeds provisioned space, or manually before peak-workload. In this project, the collections are scaled-up beginning of each month, for the duration of data ingestion (by pipeline (1) and (2)) and later (mid-late month) for feature extraction and modelling (7, 8). Outside of these periods, the collections are scaled-down, to minimize running cost.
- **Complete indexing**: In striking contrast with traditional SQL databases, Cosmos DB indexes all the data properties inserted, by default. This is paramount for feature selection and engineering performance, even on obscure nested features of the social post, not foreseen by first experiments. New hypothesis do not require updating schema or re-indexing content.
- **Distribution**: Cosmos DB partitions and their replicas are distributed geographically, with no documented limit for the number of read-enabled replicas. This means every Spark worker node can work with a dedicated partition individually, effectively maximizing the feature extraction and experimentation throughput.
- **Multimodel support**: there are four different data models and associated APIs available at the time of creating a collection (data store). While this project relies mostly on DocumentDB model with SQL API, for raw JSON storage and analysis, the Graph model is a noteworthy and powerful alternative for future modelling of networked data.

The choice of Cosmos DB as a sink for UGC pipelines (1) and (2) and source for analytics (7), allowed feature extraction at a rate of 44.000 posts per second. This is roughly 5x the average velocity delivered by the full Twitter Firehose. There are however, two significant caveats to use of Cosmos DB. Any update

to an existing document requires re-indexing (i.e., delete old, then insert new document), and the cost of maintaining an on-line database of millions of documents, scaled out to accept high volume and velocity of privacy requests, is non-negligible for a single Ph.D. project.

3.4.2 Auxiliary Store: Azure Data Lake

Azure Data Lake Generation 2 is a new Big Data storage solution from Microsoft, backed by Azure Blob Storage and regular HDD drives, offering a scalable yet cost-effective alternative for Cosmos DB. A substantial amount of non-sensitive data is produced by pipelines (5), (6) and (7 - described later), which does not require the application of compliance signal simultaneously with analysis, including:

- Content Engagement Totals from Twitter are anonymous and publicly available
- Instagram User Network Totals are publicly available for web crawlers around the world, with no compliance signal available
- Sentiment Analysis results are anonymous and do not enable reconstruction of the input text
- Image Recognition results are anonymous and do not enable reconstruction of the input images

The above-listed data does not require scalable exposure to privacy signal; however, it is, in another contrast to UGC, subject to change. Content Engagement and User Network Totals are expected to change in time. The same can be said about the Sentiment Analysis and Image Recognition methods implemented during this project. For the above reasons, Azure Data Lake Gen. 2. is chosen as the secondary storage, to minimize the cost per operation while still offering an acceptable read performance for analytics. Table 3.1 offers a summary of

Data type	Source	Purpose	Destination	Indexing	Partitioning	Consistency
(2) Historical tweets (raw)	Historical PowerTrack	model input	Amazon S3	n/a	date	strict
(2) Historical tweets (raw)	Amazon S3	model input	Cosmos DB	full, lazy	author.id	eventual
(3) Compliance requests (delete)	Compliance Firehose	compliance	Cosmos DB	n/a	author.id	eventual
(1) Instagram posts (raw)	Data Sift	model input	Cosmos DB	full, lazy	activity.id	eventual
(1) Instagram Engagement Totals	Data Sift	supervision	Cosmos DB	full, lazy	activity.id	eventual
(5) Instagram User Totals	World Wide Web	model input	Data Lake G2	user.id	none	strict
(6) Twitter Engagement Totals	Engagement Totals	supervision	Data Lake G2	activity.id	none	strict
(7) Sentiment predictions	Cosmos DB	model input	Data Lake G2	activity.id	dataset.id	strict
(7) Image recognition predictions	Cosmos DB	model input	Data Lake G2	activity.id	dataset.id	strict

Table 3.1: Data flow, partitioning and serving summary

storage solutions across project pipelines. Eventual consistency and lazy indexing are chosen for all Cosmos collections to minimize the cost and latency of insert operations. Partitioning on the ID of the author is crucial for scalability of Compliance Replay, ensuring that user-delete requests are not fanned-out to all Cosmos partitions, but served by a single partition each.

3.5 User Generated Content collection

User Generated Content (UGC) analyzed in [KL19, KH20] is collected in batches of JSON formatted activities from external API's offered by Twitter and DataSift. The batches exhibit high variability in volume, velocity and variety, throughout the project. The goal of the tasks described in this section, is to enable a uniform access to all of these activities via SQL API, for data science and privacy compliance requests simultaneously. Data collection pipelines (1) and (2) are responsible for retrieval, enrichment and indexing of UGC. Below motivates the choice of content and enrichment APIs, as the collection pipelines' dependencies:

- **Twitter's PowerTrack (PT) Filtering**: originally developed by GNIP, provides users with the ability to filter the full Twitter Firehose only to receive the activities of interest. Introduces a wide variety of filtering operators to match tweets based on user attributes, content attachments, geo-location, and many others. The use of PowerTrack filtering provides a significant opportunity for the project, to limit data noise already at the source, before the collection. Depending on the modelling task, this could mean focusing the collection only on original Tweets (i.e., excluding retweets) or languages supported by Sentiment Analysis. Any noise reduction translates directly to the reduction of collection and compliance workloads, to be discussed in the following sections.

- **Twitter's Historical Power Track (HPT)** is an API launched by GNIP in July 2012, with the aim to offer PT filtering capabilities for the entire Twitter Archive. HPT is designed for extracting tweet volumes at scale. Every public tweet ever posted can be retrieved via HPT. [TBDH19] argues that the relatively new field of computational social science is still plagued by anecdotal-based arguments, where most analyses focus on small samples, covering only a small period, leading to significant bias in attempts to explain all events and making future predictions. Relying only on short time-frame samples or keyword-based crawling can produce a large dataset full of noise and irrelevant [BAK17] data. The opportunity HPT creates for science is substantial. If the HPT job is accepted, every single Tweet posted during the period of interest expressed

within a PT filter is examined for a match. An HPT job that covers 14 months results in filtering of nearly 300 billion Tweets, often in less than 24 hours. Figure 3.3 illustrates the volume per month distribution of an example dataset collected for this study. Building a similar dataset by sampling Twitter's Firehose, which dominates prior work, would have taken 14 months minimum.

- **Data Sift's Activity Streams**: user-generated content from many networks other than Twitter is made available for analysis by Data Sift. This data provider offers dedicated live activity streams per network of interest. A single stream delivers batches of UGC and (optionally) engagement counts for content within filters. For this project, activity filters have been deployed to match both available UGC and engagement counts from Instagram. After a month of collection within this project, Instagram support has been discontinued by DataSift.

For this project, multiple years of Twitter History have been filtered using PT rules and operators, to minimize data points considered noise and focus the collection on relevant data only. The filters have been applied retroactively to extended time periods, to increase the size of the population represented by the datasets, and maximize the generalizability of any resulting models. The Instagram part of this study is based on an unfiltered stream ingested by Microsoft over a period of 1 month. The 3Vs of mostly privacy-sensitive big data exposed

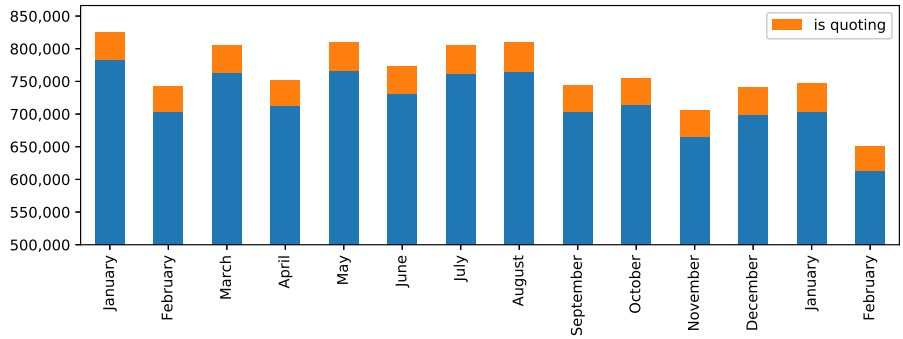


Figure 3.3: Extended time-frame sampling with Historical PowerTrack API

by Twitter and DataSift motivate the implementation of custom distributed data ingestion pipelines and a storage solution, with one ultimate goal in mind: to maximize signal-to-noise ratio for privacy-respecting data science at scale. Figure 3.4 illustrates a high-level design of the ingestion pipelines, developed during this project. Pipeline (1) is responsible for enrichment, registration, and indexing of Tweets after collection from the HPT storage (Amazon S3).

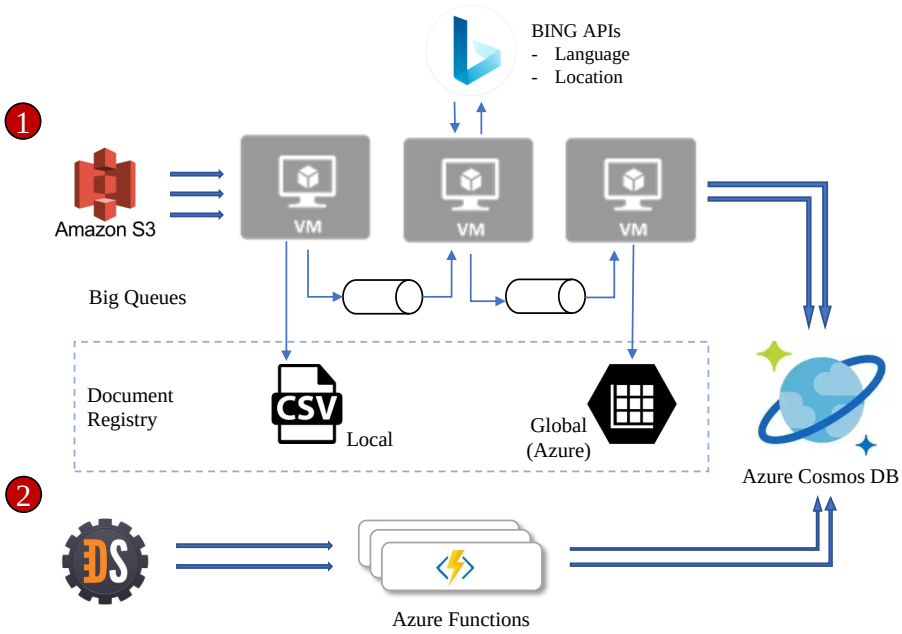


Figure 3.4: Data collection pipelines collect and store JSON activities from Twitter (1) and Instagram (2) at high volume, velocity and variety

Pipeline (2) is responsible for indexing entire Instagram traffic consumed by Microsoft in a cost-effective way. The following page summarizes key design choices taken with these responsibilities in mind.

- **High-concurrency entry buffering**: The moment HPT finishes filtering Twitter's history for content matching the requested PT rules, the resulting volume is delivered to Amazon S3 storage, in the form of GZIP-ped batches of JSON activities, to be collected within 15 days. Retrieval of the batches from across the Atlantic, to a data center in Europe, is highly parallelizable due to date-based partitioning at the source, allowing peak throughput of 400Mb/s before decompression. Once in Europe, each batch is decompressed. Every activity within the batch is individually registered in a CSV-based on-premises registry, before enqueuing for enrichment by the second stage processors.
- **Persistent inter-process communication**: in order to maintain a high throughput across the stages of ingestion, it is crucial to limit any unnecessary overhead. In cloud computing, the golden standard of inter-

process communication, motivated primarily by reliability and scalability demands, is to rely on remote queues like those offered by the Azure Service Bus. In the case of HPT ingestion, the network communication overhead inherent to remote queuing systems was the first thing to avoid. For inter-process communication, the high volume ingestion pipeline relies on a persistent memory-mapped message queuing solution called BigQueue. The following section provides a design summary.

- **Bing Translation and Location APIs**: The ingestion pipeline provides an excellent opportunity for addressing any data issues, including impurity, missing, or denormalized fields. Most of Twitter activities are delivered with language detected upfront. The rest is subject to additional language recognition using Microsoft Translation Text API [Mic19b]. A minority of incoming content contains some form of location information. The consistency and granularity of the incoming location data vary. While the majority of the posts include none, some specify a country code (e.g., 'US'), city, or a vernacular region (e.g., "Big Apple"). Bing Maps Location API [Mic19a] is called to find the GEO coordinate of the center, for all such cases. Data Science tasks dependant on these coordinates are out of scope for this Ph.D. project, and collected towards future contributions.

- **High-availability Azure Functions**: In contrast to HPT collection, the ingestion of the incoming DataSift streams has a very low tolerance for downtime, measured in minutes, after which data loss occurs. In order to keep up with inbound streams of variable velocity, while minimizing cost, the indexing implementation needs to support rapid scale-out. The golden standard for scale-out today is offered by microservices. [GWZW16]. Part of this project, the Instagram indexing process is developed as a lightweight Azure Function [MB17], with number of instances rapidly adjusted depending on the incoming traffic.

- **Distributed, adaptive indexing**: The common final goal of both collection pipelines is to ensure every document is stored at the destination, in a way that maximizes the performance of consumption scenarios, specifically the serving of feature extraction and compliance requests. This process of indexing cannot assume the volume of incoming data, the number of other operations performed on the storage simultaneously, and consequently, the throughput of the destination. The HPT indexer removes a batch of documents from the back of incoming BigQueue, only upon confirmation from the Cosmos DB management node. The Instagram dedicated Azure Functions send a confirmation response HTTP 20x to DataSift, only after Cosmos DB confirms the received batch. In case of destination overload, the thread responsible for a particular batch attempts a retry only after a period requested by the destination, specified part of the HTTP 429 error message. This adaptive implementation maximizes the throughput of the

pipeline while minimizing data loss. With Cosmos DB collection indexing set to lazy, the Twitter indexer achieves peak throughput of 8000 tweets/s, matching the average velocity of the full Firehose.

3.5.1 Document Registries: Global and Local

A substantial amount of data processed in this project is subject to privacy requests and regulations. The best-known way to adhere to privacy regulations is to respect each privacy request received. Considering the mean rate of 2,000 (peak of 8,000) compliance requests per second observed from Twitter Compliance Firehose in 2019, serving these requests on time becomes a big data engineering problem, fast. The work required to cross-reference all the incoming requests with all the HPT content stored in Cosmos DB, to retrieve and delete, would dramatically impact the scalability and cost of the solution, and likely prevent this project from completion. This critical concern motivates introduction of two document registries. Below summarizes the primary and secondary use cases of each:

- **Local CSV-file based registry**: stores every unique document ID and author ID processed during this project. The primary use case is to facilitate the pre-filtering of the Compliance Firehose Replay. Early filtering drastically reduces the volume of requests forwarded to Cosmos DB. Furthermore, the document IDs also inform a focused retrieval of Content Engagement Totals. Both of these processes require dedicated pipelines, described in the following sections.

- **Global Azure-based registry**, implemented as a Service Bus Table, for high-availability across projects. This registry, aside from the document and author ID's stores the precise location(s) of sensitive content, and informs any retrieve/update/delete operations. This purpose renders it a critical dependency of the Compliance Processor, part of the Compliance Pipeline described in the following section.

3.5.2 Persistent inter-process communication at scale

The diversity of type and computational cost of big data processing tasks in this project validates the encapsulation of responsibility in dedicated processes with assigned hardware. The isolation of sequential responsibilities demands a scalable and reliable solution for inter-process communication. Scalability and

reliability goals are frequently at odds with each other. The relatively fast inter-process communication facilitated by the operating system (e.g., named pipes) is vulnerable to frequent unplanned shutdowns or synchronization issues (e.g., single file shared across processes). In cloud computing, services like Azure Service Bus, offer message queuing designed to address both the synchronization and fault tolerance concerns. The elaborate locking and redundancy architecture, along with the network communication overhead, can severely impact the overall throughput and scalability of the solution. In this project, high volume pipelines rely on neither. The advantages of a memory-mapped persistent

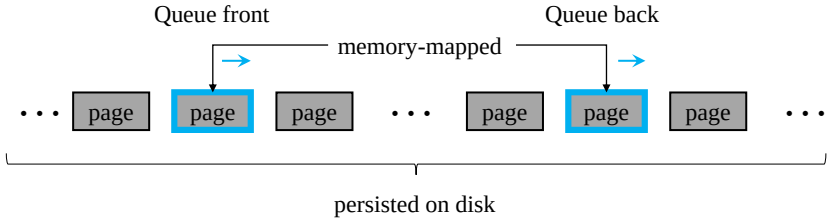


Figure 3.5: BigQueue: persistent memory-mapped queue for inter-process communication at scale

queue, offered by the BigQueue library [Lea16] in the context of this project, are summarized as follows:

- **Performance**: enqueue and deque operations are performed on the front and rear of the queue, as illustrated by figure 3.5, which results in IO performance close to O(1) and comparable with direct memory access. Memory-disk mapping dramatically increases the throughput of each pipeline stage, relative to remote queues.
- **Scalability**: the total size of the queue is limited only by the space available on the provisioned disk. Such flexibility is crucial during retrieval of high volume HPT batches from Amazon S3, which quickly exceed the size of memory or system hard drives of the executor machines, both also subject to data loss during an unexpected shutdown.
- **Reliability**: every page of the queue is persisted on the physical hard drive provisioned for the pipeline's machine. The operating system is responsible for persisting the produced messages, even if the pipeline process

crashes. Built-in synchronization ensures that multiple pipeline processor can concurrently enqueue and dequeue without data corruption. Such a safeguard makes it possible to run additional instances of a particular processor when mitigating throughput bottlenecks.

3.6 Privacy compliance pipelines

Every single User Generated Content (UGC) analyzed in [KL19, KH20, RKH20], has been publicly available at the time of data collection. Exactly how much of it remains public, can change rapidly afterward. Account removal, suspension, or deleting of a single tweet render all related content unavailable for compliant analysis. Users exercise their right to be forgotten at an unprecedented rate [BBC+19, KL19]. Many of them have their accounts and content removed automatically, part of Twitter's crackdown on fake accounts and foreign political interference [SRN+18b]. In this project, instead of anonymizing the datasets, sensitive or private information is eliminated from storage and future analysis as soon as the request from the user is processed by the social media platform. In 2018 there has been an average of 2,000 (with peaks of 8,000) such delete requests delivered by Twitter's Compliance Firehose. Ingestion, archiving, filtering, and application of these requests are the responsibilities of Compliance Pipelines (3) and (4), with the architecture summarized by figure 3.6.

3.6.1 Online Compliance

The volume of incoming privacy requests is delivered to Microsoft by Twitter's Compliance Firehose via multiple redundant live streams and consumed by the Online Compliance pipeline (3). The responsibility of this pipeline is to ensure uninterrupted consumption of the stream, application, and archivization of all incoming requests in the online Compliance Archive hosted by Azure. The choice of Blob Storage for the archive is motivated primarily with cost savings due to the large volume of the incoming requests. The first stage processors perform archivization after de-duplication of redundant data streams, before passing the requests to downstream processors via Azure Service Bus. The second stage processors are responsible for filtering out the incoming requests concerning content or users unknown to Microsoft. The filtering is informed by the global Document Registry, which also contains all the location(s) of sensitive content. This information is propagated along with the request to the final stage processors, responsible for complete removal of UGC or an author, from the storage. The partitioning design of Cosmos DB collections ensures that any

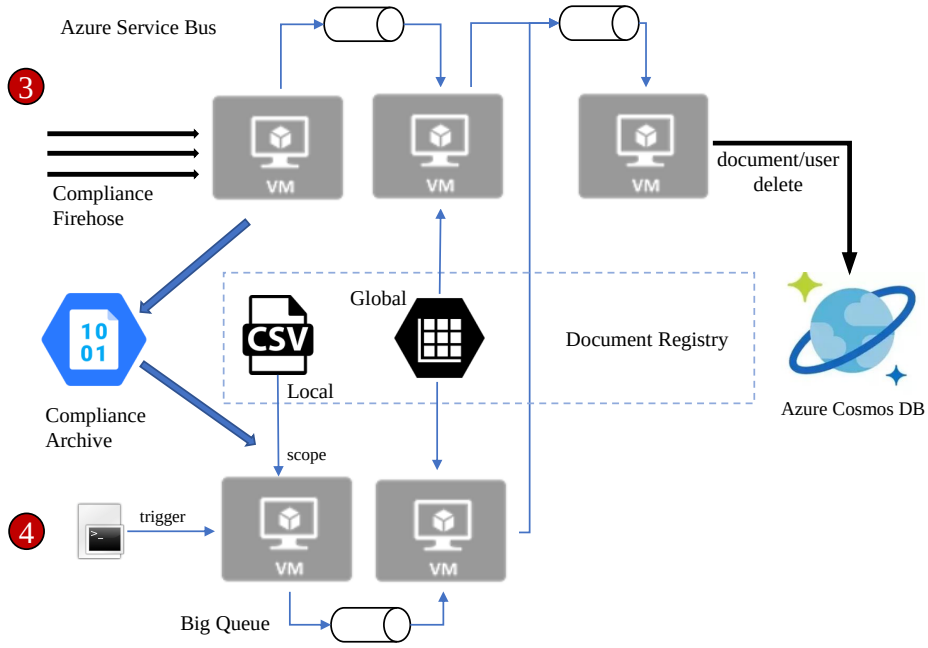


Figure 3.6: Privacy Compliance pipelines: Online (3) and Replay (4) ensure every applicable privacy request is reflected by the central storage solution

user delete request can be served by a single Cosmos DB node (partition) only, which ensures maximum end-to-end throughput of the compliance infrastructure while minimizing the impact on modelling and specifically, feature extraction performance.

3.6.2 Compliance Replay

One week of processing by the collection pipelines creates a compliance debt of at least 1.2 billion requests. All of them have to be examined for a match with owned UGC, and if necessary, applied to any known copy of the no-longer public UGC, all within the grace period. This is ensured by the Compliance Replay pipeline (4) triggered ad-hoc after successful data indexing by the collection pipeline. The process begins with retrieval of all compliance requests issued in the period of transit, from the Microsoft controlled Compliance Archive, popu-

lated live by the Online Compliance pipeline (3). The volume of these requests can easily overwhelm downstream infrastructure and impact the performance of the dependent analytics and modelling tasks. With this concern in mind, entry filtering is implemented, based on the local Document Registry, populated during data ingestion. The entire list of registered, unique document IDs is read into each processor's RAM to facilitate the filtering of all incoming requests at high velocity. The vast majority of non-applicable requests (regarding UGC out of Microsoft control) is discarded, dramatically limiting the workload for downstream infrastructure. Occasional hash-collisions lead to false positives propagated to the next processor, which filters them out based on the global registry, or in case of a match, forwards the request to the pipeline (3), for execution.

3.7 Content and User Engagement Signals

The social engagement modelling tasks described by Chapter 2 and the scientific papers enclosed, all depend on the engagement signals unavailable at the time of UGC collection [KL19, KH20]. Below three multidimensional signals require dedicated collection effort each:

$$\epsilon^{content}_{Twitter} = [\epsilon_{\text{retweets}}, \epsilon_{\text{replies}}, \epsilon_{\text{favorites}}] \tag{3.1}$$

Twitter content engagement signal is publicly available as the number of retweets, replies, and favorites received by a tweet, or a 3-dimensional vector 3.1.

$$\epsilon^{content}_{Instagram} = [\epsilon_{\text{likes}}, \epsilon_{\text{comments}}] \tag{3.2}$$

Instagram content engagement is publicly available as the number of likes and comments received by a post, or a 2-dimensional vector 3.2.

$$\epsilon^{user}_{Instagram} = [\epsilon_{\text{followers}}, \epsilon_{\text{following}}, \epsilon_{\text{posts}}] \tag{3.3}$$

Finally, the Instagram author engagement signal is publicly available, as the number of followers, accounts followed, and the count of UGC owned by the author, expressed by the 3-dimensional vector 3.3.

3.7.1 Twitter Engagement Totals API

There are two ways of collecting $\epsilon_{Twitter}^{content}$, which dominate prior work on Twitter engagement modelling:

- Web crawling: for every tweet, the WWW representation is retrieved, and the components of the vector are parsed from the raw text. This method represents the present counts at the time of crawling but does not include engagement actions removed by, e.g., privacy requests, affecting the accuracy of ML supervision. Furthermore, the significant communication overhead involved in crawling and vulnerability to automatic anti-crawling measures by target sites, drastically reduce the scalability of the approach.
- Activity counting: UGC collection, in this case, is not limited to unique content, but also replies and retweets, to be counted. This approach requires full Firehose access, introduces substantial overhead in collection effort and a proportional increase in compliance responsibility.

Part of this study, an alternative way of sourcing engagement signal is developed, based on Twitter Engagement API, and summarized by Figure 3.7.

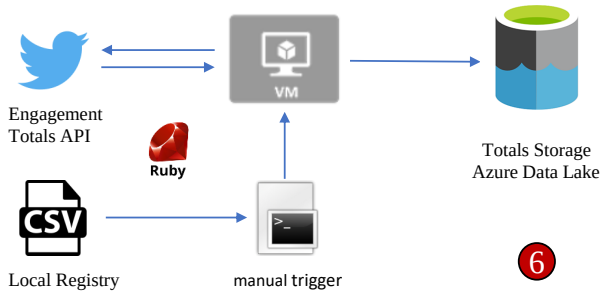


Figure 3.7: Content engagement signal is extracted using Twitter's Engagement Totals API

Twitter Engagement Totals pipeline (6) is executed as soon as the collection by pipeline (2) is completed. The process is designed to retrieve the vector 3.1 for all Tweets of interest, identified by the local registry instance, populated during collection. The requests are issued in batches with 2 seconds of delay between, to maximize the rate limits of the API. Finally, the signal is stored in the target

Data Lake Gen. 2, exposed to requests during the modelling stage (7). This solution secures the total counts of engagement actions (retweets, replies, or favorites) ever registered for the specified UGC (even if removed later), thereby maximizing the accuracy of supervision for predictive analytics.

3.7.2 Instagram User and Engagement Totals

The Instagram content engagement totals described by vector 3.2 are delivered by DataSift in the form of JSON activities, in a similar way as the UGC. This enables the reuse of pipeline (1) for consumption, before processing and conflation at a later stage (7). Unfortunately, contrary to Twitter's, Instagram activities do not arrive with user counts embedded. The number of followers, in particular, is well recognized by prior work, as one of the strongest predictors of engagement. This motivates an additional effort to extract 3.3, summarized by Figure 3.8. While all the user counts of interest are publicly available, any

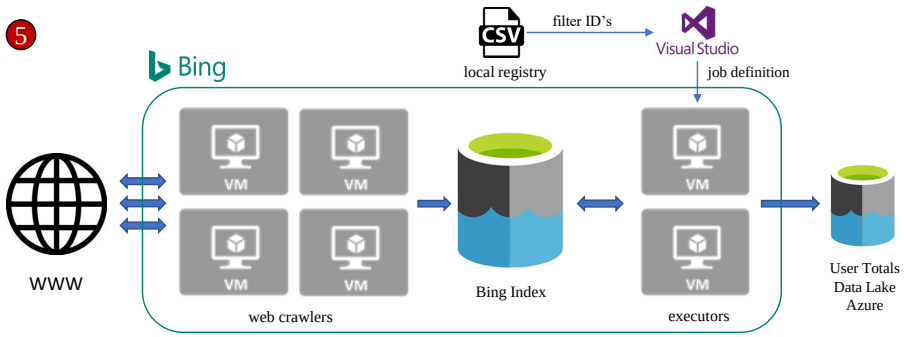


Figure 3.8: User Totals crawling infrastructure powered by Microsoft Bing

custom attempts to crawl the websites of Instagram have triggered automated rate-limiting or blacklisting of the originating IP addresses. There are, however, many services that have been whitelisted by Instagram and exempt from the rate limits. The publicly available robots.txt file (Instagram), reveals one owned by Microsoft, called the Bingbot. This project relies on the Microsoft Bing web crawlers to retrieve volumes of vectors 3.3 for Instagram engagement analysis. The overview offered by Figure 3.8 illustrates two parallel data flows: a high-concurrency recurring flow between Bing's crawlers indexing the public websites, and the ad-hoc flow implemented for this project, to retrieve only the features of interest, from selected Instagram pages.

3.8 Feature Extraction at Scale

The goal of previously described collection and compliance pipelines was to render most of the velocity, variety, and sensitivity challenges transparent for the data science workloads. The goal of the architecture (7) is to prepare the richest possible representation of UGC for privacy-compliant engagement modelling at scale. In machine learning, the importance of quality feature engineering needs little introduction. This is arguably the most computationally expensive stage of the project. It has a direct impact on the final performance, robustness, and scalability of the models, as measured in Chapter 2.

Every Twitter and Instagram UGC analyzed in this study is collected in the form of JSON activities, and stored in partitioned, distributed collections of Azure Cosmos DB. Each of these activities offers a nested, multidimensional representation of the original post, referred to as the RAW content, and well documented by the data providers [Twi19, Dat19]. Only a minority of the RAW content dimension offers a signal relevant for engagement modelling. The choice of these dimensions has been summarized in Chapter 2, validated by prior work and ablation studies [KL19, KH20]. This section describes the process implemented to extract and maximize the signal-to-noise ratio, from the RAW UGC at scale.

Figure 3.9 illustrates the scope and distribution of this process. The synergy of Data Science and Big Data Engineering, in this thesis, begins with Cosmos DB [RR18] and Apache Spark [ZXW+16] integration. The guiding assumption behind all the technical choices described below is that the end model quality depends directly on the number of the hypothesis tested (i.e., mistakes learned from), under time constraints explained in section 3.3. Consequently, the motivation of the architecture 3.9 is to minimize the idle periods in the project (e.g., waiting for processing results) and to maximize the number of experiments that can be conducted in a fixed amount of time. This project relies on the Apache Spark cluster for all data science tasks described in Chapter 2. Apache Spark is well-suited for the development of large-scale machine learning applications due to an adaptation of the Map-Reduce approach to iterative computations on distributed in-memory data structures [MBY+16]. This approach is often credited, for the Spark's ascend to the de-facto industrial standard for Big Data analytics, and the rapid growth of Ali Ghodsi's Databricks.

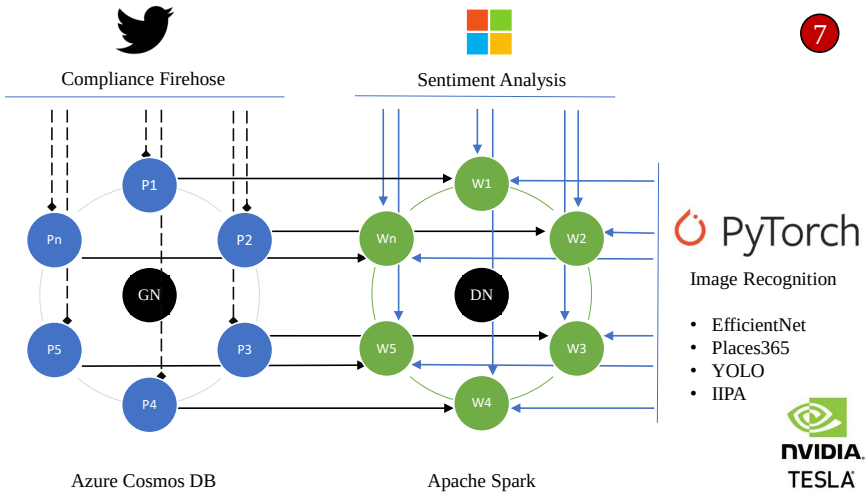


Figure 3.9: Cosmos DB and Apache Spark integration: Apache Spark Driver Node (DN) distributes feature engineering workload to worker nodes Wx. These execute feature extraction from the selected RAW dimensions, served by dedicated Cosmos DB partitions Px.

3.8.1 Resilient Distributed Datasets

RDD is an abstraction offered by Spark to boost the efficiency of a wide range of iterative algorithms and data mining. RDDs allow Spark to outperform existing models (e.g., Hadoop) by up to 100x in multi-pass analytics [GA15]. In this project, the UGC representation, created from selected RAW dimensions, extracted via SQL API is no longer personally identifiable, however, it remains in-memory with no persistent representation created by Spark. Restricting the analysis to in-memory RDDs allows maximizing the performance of all the tasks described in Chapter 2, but also prevents fragmentation of any sensitive data outside of the central Cosmos DB, exposed to user privacy requests [KL19].

3.8.2 Distributed feature collection

Many of the dimensions of the RAW content (e.g., handle, biography or type of browser used) are irrelevant for this project. Use of Cosmos DB SQL API allows focusing feature extraction only on selected dimensions. This helps limit the communication overhead with the collections or transmission of any privacy-sensitive or otherwise personally identifiable information (PII). For example, where RAW content would contain the exact handles of users mentioned in the tweet, the array_length operation will only return the number of accounts mentioned. The SQL API is also useful in focusing the requests on a specific period, topic, or origin of the UGC collected. The distribution of workload between the Spark cluster and partitioned Cosmos DB collections, enable feature extraction at rates exceeding 65,000 tweets per second, thereby accelerating every dependant experiment.

3.8.3 GPU accelerated feature extraction

Predictive analysis towards the extraction of features from natural language and vision is far more complicated, and historically motivates most of the activity in the field of deep learning. Below summarizes the project dependency on some of the most promising and complex deep neural architectures available today. If Deep Residual Networks are credited for achieving human-level performance in image recognition, the recent Bidirectional Encoder Representations from Transformers represent such a breakthrough for natural language processing (NLP). The choice of below deep models to enhance UGC feature representation was motivated in Chapter 2 [RKH20]. Their architectures are well documented in peer-reviewed journals, after multiple years of research and development.

- **Sentiment Analysis**: State-of-the-art models in sentiment analysis today, including the one delivered by Microsoft Cognitive Services [Mic17] relies on a recent language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. The pre-trained BERT model is fine-tuned with at least one additional output layer, to advance a wide range of tasks, including sentiment analysis. BERT obtains new state-of-the-art results on eleven natural language processing tasks, including GLUE, MultiNLI, SQuAD v1.1, and SQuAD v2.0. This advancement is attributed to multiple layers of attention (12 or 24 depending on the model), and multiple attention "heads" in every layer (12 or 16). Since model weights are not shared between layers, a single BERT model effectively has up to 384 different attention mechanisms [DCL+18]

- **EfficientNet**: Convolutional Neural Networks (ConvNets) are the class of DNN most commonly applied to analyzing image data, and commonly developed to maximize the available resources. EfficientNets family of architectures [TL19b] is a result of a neural architecture search to design a new baseline network, which achieves much better accuracy and efficiency than previous ConvNets. The EfficientNet-B7 achieves SOTA on ImageNet while being 8.4x smaller and 6.1x faster on inference than the best existing ConvNet. The EfficientNets potential in transfer learning is proven on CIFAR-100, Flowers, and three other datasets, with an order of magnitude fewer parameters than previous best ConvNets.

- **YOLO: You Only Look Once**: In the field of computer vision, object detection is particularly challenging, as it involves a combination of object classification and object localization within a scene. DNNs continue demonstrating superior object detection performance compared to other approaches. YOLO is one SOTA approach in DNN-based object detection in terms of both speed and accuracy. The real-time performance, however, is not possible without a powerful GPU [RDGF16, SCLW17]

- **IIPA: Intrinsic Image Popularity Assessment**: This pre-trained DNN has been developed to predict the potential of a social image to go viral on the Internet. It is optimized for ranking consistency with millions of popularity-assessed image pairs. The authors claim human-level performance on Instagram. It is based on ResNet-50 [HZRS16a] DNN architecture, and requires approximately one day to train with NVIDIA Tesla V100 GPU [DMW19]

- **WideResNet-18: Places365** is trained on an image database for deep scene understanding [ZLK+18b], released part of a challenge to advance at the task of visual scene recognition. The pre-trained model used in this project is based on WideResNet-18 architecture. Deep ResNet's are able to scale up to thousands of layers and still show performance improvements. However, training very deep residual networks have a problem of diminishing feature reuse, making these networks very slow to train. The novelty of WideResNets lies in decreasing the depth and increasing the width of residual networks. [ZK16] show the advantages of this approach over their commonly used thin and deep counterparts. They demonstrate that even a simple 16-layer WideResNet outperforms all previous deep residual networks, achieving new SOTA on CIFAR, SVHN, COCO, and significant improvements on ImageNet.

The above summarized DNN dependencies taken in this project were until recently prohibitive at scale. Given a single machine with a 6-core Intel i7 processor, the sentiment analysis of all Twitter UGC in this project, would require six

months. Extracting image features described in Chapter 2, from all the Instagram UGC collected, would have taken five years. The most computationally intensive part of the project, motivates a GPU accelerated Apache Spark cluster. Every DNN model mentioned is CPU-bound OOB (out-of-the-box). Each of them has been carefully ported to CUDA part of this project and deployed to Spark worker nodes with NVIDIA Tesla V100 GPUs on-board. The distributed architecture proposed in Figure 3.9 allows a speed-up of 70x, and thus enables transfer learning in modelling social engagement at scale.

3.9 Engagement modelling at Scale

The diverse set of signals collected and processed by the pipelines (1-7) finally positions this project for privacy-respecting social engagement modelling at scale. Chapter 2 has been dedicated to the results of this phase alone. The purpose of this section is to describe key design choices in maximizing the performance and accuracy of all dependant tasks. With the velocity, variety, and sensitivity concerns encapsulated by the previously described workloads, the guiding motivation remains to maximize the predictive performance of resulting models, via maximizing the speed of analysis. The analytics architecture proposed for the project's modelling phase is summarized in Figure 3.10. It is rarely discussed how many mistakes and failed attempts are necessary to produce one strong and robust model. Each of these mistakes offers some educational value for a researcher (especially at an early stage of a scientific career). Therefore the analytics solution, intuitively, should maximize the number of experiments that can be done within a fixed period of time. The architecture should secure

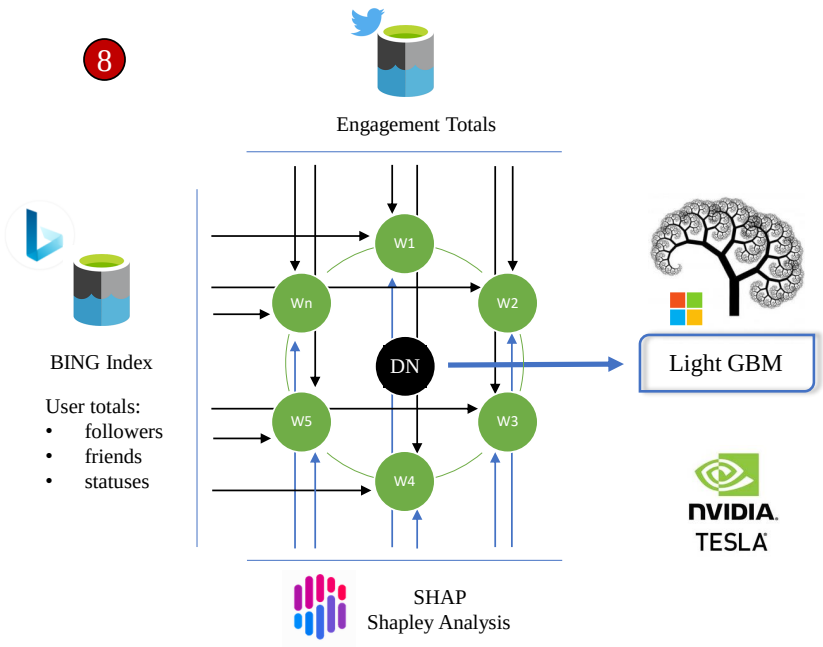


Figure 3.10: modelling environment, powered by Apache Spark and NVidia, facilitates signal conflation, exploratory and predictive analysis and finally, explainable AI.

the researchers' focus on the scientific questions while minimizing the many distracting delays caused by the method and resource constraints. This is perhaps a common opportunity for the industry-academic partnerships.

3.9.1 GPU accelerated Gradient Boosting

The engagement models are trained using the LightGBM [KMW+17] framework from Microsoft. Prior work on multiple public datasets shows that the framework's features including Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) can accelerate the training process by over 20 times at negligible impact to accuracy [KMW+17]. Finally, LightGBM implements both: [Wal58] for handling of categorical features, and histogram-based algorithm to approximately find the best splits, which is highly scalable on GPUs [ZSH17]. The relatively low dimensionality of engagement prediction tasks on Twitter does not require GPU acceleration. As described in Chapter 2, the modelling tasks steered away from computationally expensive black-box approaches like DNN's and relied on less than 30 input features. This is many orders of magnitude less than the dimensionality of ResNet or BERT architectures' input data. However, the histogram-based algorithm benefits dramatically from GPU acceleration [ZSH17]. The speedup allows traversal of a substantially larger hyper-parameter space during cross-validation, outlined by Algorithm 1.

The amount of data used for training any of the engagement models in this project does not require Apache Spark distribution at this stage. The amount of noise removed from RAW content by pipeline (7), reduces (or summarizes) an incoming HPT dataset T2018 of 93 GB RAW sensitive data, to less than 400MB of highly relevant anonymous signals. However, the highly complex Shapley Analysis [LEL18], which consists of many independent iterations, scales near-linearly with the number of cores available. Calculating explanations for all the Engagement predictions on a 96-core cluster showed a speedup of 93.3x.

Algorithm 1 describes an example workflow of a modelling task executed end-to-end in the context of the Spark Driver Node (DN) with an NVIDIA Tesla V100 GPU on board. DN is the single source of control in the architecture 3.10. The Virality model trained by the DN outperformed previous state-of-the-art models by over 30% on multiple benchmark datasets (Chapter 2). The training of the final model was completed in 4 minutes. The report on the results was awarded Best Paper [KL19].

Algorithm 1: Twitter Engagement 2018 model training with nested K-fold cross-validation and hyper-parameter tuning

Result: Gradient Boosted Regression Tree
// Assume C is HPT CosmosDB collection, L is Totals Data Lake
$H \leftarrow hyperparameter\ combinations$
$RDD_{UGC} \leftarrow extract_{(7)}(C_{T2018})$
$RDD_{Totals} \leftarrow read(L_{Totals})$
$RDD_Y \leftarrow compute\_engagement(RDD_{Totals})$
$RDD_{XY} \leftarrow inner\_join(RDD_{UGC}, RDD_Y)$
for $i \leftarrow 1$ ***to*** K **do**
  $XY_{Test} \leftarrow collect(RDD_{XY})_i$
  for h ***in*** H **do**
    for $j \leftarrow 1$ ***to*** $K-1$ **do**
      $XY_{Validation} \leftarrow collect(RDD_{XY})_j$
      $XY_{Train} \leftarrow collect(RDD_{XY})_{K-2}$
      $GBRT \leftarrow train(XY_{Train}, h)$
      $Y \leftarrow predict(GBRT, X_{Validation})$
      $\rho_j \leftarrow spearmanr(Y, Y_{Validation})$
    end
    Calculate avg. perf. over K-2 folds for hyper-param. combination h
  end
  $XY_{Train} \leftarrow collect(RDD_{XY})_{K-1}$
  $GBRT \leftarrow train(XY_{Train}, h_{best})$
  $Y \leftarrow predict(GBRT, X_{Test})$
  $\rho_i \leftarrow spearmanr(Y, Y_{Test})$
end
Calculate average performance over K folds

3.10 Engagement Analysis in Production

The delivery of results from the modelling phase of this project, cannot and does not end with the academic avenues. The industrial nature of this Ph.D. collaboration requires an ongoing contribution to the business. The mission of Microsoft Corporation is to empower every individual and every organization on the planet to achieve more (Satya Nadella, 2015). The primary focus of Business Applications Group within Microsoft is to contribute to this vision with intelligent cloud services. My attempt to align the project with these goals is faced with additional challenges across disciplines. Two of the technical ones are summarized below. This section describes an engineering solution to both.

- **Volatile consumption patterns at scale**: Every engagement model in this project to be integrated with Microsoft cloud services need to be able to cope with the demand generated by users around the world. The demand is characterized by the natural daily cycles across timezones or the growth in consumption, most often tracked by the Monthly Active Users (MAU) metric. The demand is further characterized by peaks due to high visibility events. Notably, increased activity in social networks, during events related to politics, climate, celebrities or football, translate directly to peak analytical workload, as soon as requested by a single customer, private or public.

- **Deteriorating model performance**: The rapid increase in MAU of online platforms like Twitter, observed over the last decade, inspired many studies on the changing usage patterns. One example is the phenomenon of adding tags to messages. Now an integral part of most online social networks (OSN), it dates back to February 2008, when introduced by Twitter [HTE10]. The changing culture of OSN users can also be illustrated by the average number of followers, followees (aka. friends), or daily activity. All of these features offer predictive value in engagement modelling, as examined in Chapter 2. Their use is expected to cause a model's performance to decline over time.

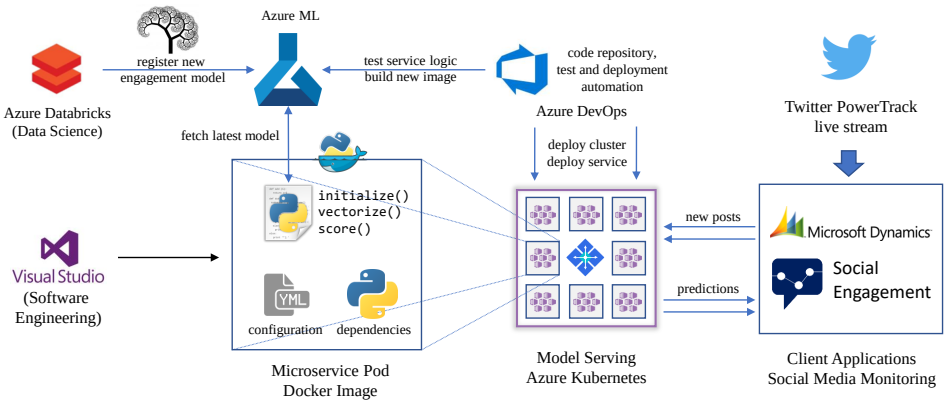


Figure 3.11: Topology of the model deployment, update and serving infrastructure, facilitating high-availability engagement analysis in production.

The architecture illustrated by Figure 3.11 summarizes the recurring tasks implemented part of this project to address the above challenges and industrialize engagement prediction at scale. Key components of the solution are described below in the context of the challenges mentioned above:

- **Custom model tuning**: The Apache Spark modelling environment described in section 3.9 and hosted by the Azure Databricks platform is responsible for the delivery of new engagement models. The training process is tuned for scalability, i.e., to balance the predictive performance with the time necessary to predict engagement for every incoming post. Specifically, only features available immediately after the post was created are considered for training. This constraint was rarely an issue in prior purely academic work. After successful cross-validation, the model is dumped to a file and uploaded to a central model registry, hosted by an Azure Machine Learning. This process is automated to facilitate model retraining, with the evolving dynamics of attention in OSNs.
- **Model input-output logic**: For every new model, custom scoring logic is developed, in Python, to acquire predictions based on batches of incoming RAW content. This process involves vectorization, i.e., extracting model input features from the incoming JSON activities (partially in line with the responsibilities of the pipeline (7)) and interpreting model's response, including inverse power transformation of the predicted score. Along with unit and integration tests, this logic is then uploaded to a GIT repository, hosted by Azure DevOps [MG18]
- **Service deployment logic**: Every external request to predict engagement will be served by a microservice pod, based on a Docker image. Docker provides an elegant abstraction of a process in times of cloud computing and for this project, an opportunity to encapsulate a model with the IO logic and all the PyPI / Anaconda dependencies. Custom logic to build and deploy these images is developed, for execution by Azure DevOps services.
- **Model serving platform**: Traditional operating system offers task scheduling via an abstraction layer from fixed on-premises compute resources (prevalent single or dual-socket Intel x86 architectures). Kubernetes [Bre15] delivers task scheduling on top of a dynamic pool of distributed compute resources in the cloud. Azure Kubernetes cluster has been provisioned to handle all incoming engagement prediction requests. The incoming requests are distributed by the Kubernetes Orchestrator across multiple instances of a microservice. Each pod represents a separate microservice based on the custom-built docker image. The number of pod instances is adjusted dynamically, based on the velocity of incoming traffic.

One critical condition to meet before going to production was the ability to serve 150 prediction requests per second. Multiple load tests of the gradient boosted microservices deployed as above, indicating a stable throughput of 300/s. The solution became globally available in April 2019 [Mic19c], serving over 8000 private and enterprise customers of Microsoft Social Engagement across the world. For the remainder of Microsoft's contract with Twitter, they were able to react to pre-viral content (including support requests) before anyone else.

Appendix A

Scalable Privacy Compliant Virality Prediction on Twitter

Scalable Privacy-Compliant Virality Prediction on Twitter*

Damian Konrad Kowalczyk[1,2][0000−0002−5612−0859]
and Jan Larsen[2][0000−0003−1880−1810]

[1] Microsoft Development Center Copenhagen, Kanalvej 7
2800 Kgs. Lyngby, Denmark dakowalc@microsoft.com
[2] DTU Compute, Matematiktorvet 303B
2800 Kgs. Lyngby, Denmark {damk,janla}@dtu.dk

Abstract. The digital town hall of Twitter becomes a preferred medium of communication for individuals and organizations across the globe. Some of them reach audiences of millions, while others struggle to get noticed. Given the impact of social media, the question remains more relevant than ever: how to model the dynamics of attention in Twitter. Researchers around the world turn to machine learning to predict the most influential tweets and authors, navigating the volume, velocity, and variety of social big data, with many compromises. In this paper, we revisit content popularity prediction on Twitter. We argue that strict alignment of data acquisition, storage and analysis algorithms is necessary to avoid the common trade-offs between scalability, accuracy and privacy compliance. We propose a new framework for the rapid acquisition of large-scale datasets, high accuracy supervisory signal and multilanguage sentiment prediction while respecting every privacy request applicable. We then apply a novel gradient boosting framework to achieve state-of-the-art results in virality ranking, already before including tweet's visual or propagation features. Our Gradient Boosted Regression Tree is the first to offer explainable, strong ranking performance on benchmark datasets. Since the analysis focused on features available early, the model is immediately applicable to incoming tweets in 18 languages.

Keywords: Twitter · virality · privacy · sentiment · explainability · scalability · popularity

1 Introduction and motivation

"The role of the social and professional networks in the spread and acceptance of innovations, knowledge, business practices, products, behavior, rumors, and memes, is a much-studied problem in social sciences, marketing and economics. Online environments like Twitter, offer an unprecedented opportunity to track such phenomena." [2]

* Supported by Microsoft Development Center Copenhagen and the Danish Innovation Fund Case No.5189-00089B

The knowledge discovery process, however, is becoming even more tangled with the arrival of social big data. 700 million tweets have been posted on the day of writing this introduction. The volume, velocity, and variety of mostly unstructured information even from a single social network are evolving at an extremely fast pace. From an engineering and data science perspective, near real-time analysis via online services and algorithms scalable in-memory are required, and demand substantial computational resources. Scientific endeavors to date offer progress toward specific subtasks of social network analysis (SNA) yet data collection and privacy compliance remain among the biggest challenges in extracting knowledge [3]. Arguably the most significant among them is privacy [34]. The social nature of nodes in these networks makes data subjective to many privacy concerns and laws. The new European General Data Protection Regulation (GDPR and ISO/IEC 27001) in force since May 25th, 2018 makes SNA and black-box approaches (like deep neural networks) more difficult to use in business, requiring the results to be retraceable (explainable) on demand [17]. In machine learning, explainable (compliant) real-time analysis is often at odds with predictive accuracy. In social popularity prediction, some of the best results today are achieved using deep neural networks, difficult to interpret [37] or data modalities time-consuming to acquire [12]. Modeling popularity relies on a precise count of responses (subject to privacy requests, i.e., retweets in virality prediction) which exposes them further. Accuracy in such studies depends on processing documents no longer available, while privacy compliance requires removing them. Ensuring accurate and explainable analysis via quality of the data and methods, while respecting user privacy, remain conflicting goals and open research issues individually. In this work we argue that significant advancement in SNA requires avoiding such trade-offs and addressing all the above issues simultaneously. We draw inspiration from multiple disciplines, to challenge state of the art in content virality prediction on Twitter. We propose a framework which to the best of our knowledge, is the first one that satisfies the properties of model preserving and privacy-compliant simultaneously. We use it to train a scalable and explainable model, and are the first to achieve strong [9] ranking performance on benchmark datasets.

2 Related work

2.1 Social big data analysis before GDPR

Social big data has become essential for various distributed services, applications, and systems [31], enabling event detection [10], sentiment analysis [11], popularity prediction [38], natural language processing, finding influential bloggers, personalized recommendation [14], online advertising, viral marketing, opinion leader detection etc. Computational and storage requirements of such applications have led to cloud scale reinvention of data storage and processing technologies. New tools are constantly emerging to replace the conventional non-effective ones, and a hybrid of techniques [20,15] is now a requirement to extract value from the social big data. [35] proposes a solution based on Hadoop technology

and a Naive Bayes classification for sentiment analysis of tweets. The sentiment analysis in performed in MapReduce layer and results stored in distributed NO-SQL data-base. [18] uses Lucene indexing with full-text searching ability on top of Hadoop for spectral clustering, to detect Twitter communities during the Hurricane Sandy disaster. In our work we pursue close alignment of data acquisition and analysis algorithms, with the strict constraints of storage and time, to accommodate both user-generated content (UGC) and privacy requests, arriving at high volume and velocity. Instead of perturbing or anonymizing the data, sensitive or deleted information is permanently eliminated from storage and subsequent analysis.

2.2 Content popularity prediction

Social network influence can be defined as the ability of a user to spread information in the network [32], with the retweet count assumed as a measure of a tweets popularity. One common challenge for content-based popularity prediction is the 140-character constraint imposed by Twitter, making it difficult to identify and extract predictive features [5]. [36] showed that carefully crafted wording of the message could help propagate the tweets better, but there's much more to UGC than the caption. [19,37] demonstrate social-oriented features were the best performers to predict image popularity on Twitter. [25] utilized textual, visual, and social cues to predict the image popularity on Flickr. [37] proposed a joint-embedding neural network combining the same cues to rival state-of-the-art methods. Recurrent and Deep Neural Networks advance feature extraction from high-dimensional unstructured data (i.e., image attachments), however due to low explainability also introduce a major drawback for critical decision-making processes (with recent advances by [33]). In this study, we prioritize explainable methods in application to structured data. [32,23,7] demonstrate relationships between the number of followers of Twitter users and their influence on information spreading. Ranking users by the number of followers is found to perform similarly to PageRank [23]. [32] models the probability to be retweeted by a power law function. [29] have used an explainable Random Forrest classifier to predict a range of the logarithm of the retweets volume. He demonstrates the predictive value of user features (e.g., count of followers), network features, and the popularity of hashtags included. [4] provide a comparison of learning methods and features, regarding retweet prediction accuracy and feature importance. They find Random Forests to achieve the best performance in binary classification of retweetability and highlight the value of author features: number of times the user is listed by other users, number of followers and the average number of tweets posted per day. [28] uses recursive partitioning trees to achieve 0.682 classification accuracy on a large topical dataset, albeit using features unavailable early (favorites count) or anymore (local publication time) challenging both scalability and reproducibility. [16] investigated the features of tweets contributing to retweetability and is the first to explore the impact of negative sentiment in diffusion of news on Twitter. We follow [16] to consider affect in our model. Substantial gains are seen when including network features extracted from the

content graph formed by retweets, or relationship graph formed by "friendships". The document level subgraphs to inform prediction are often acquired via real-time monitoring of the diffusion process. [39] predicted the popularity of a tweet through the time-series path of its retweets, using a Bayesian probabilistic model. [37] uses preconditioned recurrent neural network to model the temporal diffusion, and shows SOTA ranking performance of 0.366 on benchmark datasets. [1] used temporal evolution patterns to predict the popularity of online UGC. [8] use temporal and structural features to predict the cascades of photo shares on Facebook. [41] model the retweeting cascades as a self-exciting point process. [12] argues that determining the topic of interest of a user based on his past tweets might boost predictive accuracy. [30] studied retweet network propagation trends using conditional random fields, demonstrating gains in accuracy when considering social relationships and retweet history. Access to subgraphs on the author or even document level is however strictly limited by social networks, thus leveraging tweets (early) performance, authors relationships, preferences or retweet history is prohibitive for a scalable, near real-time prediction on a single tweet.

In this study we seek to maximize virality ranking performance. We follow [37] to approach the problem as Poisson regression, and [16] to consider tweet sentiment in prediction. However, in the contrast to prior work, we don't sacrifice scalability or privacy compliance, nor rely on available retweet count for ground truth.

3 Solution overview

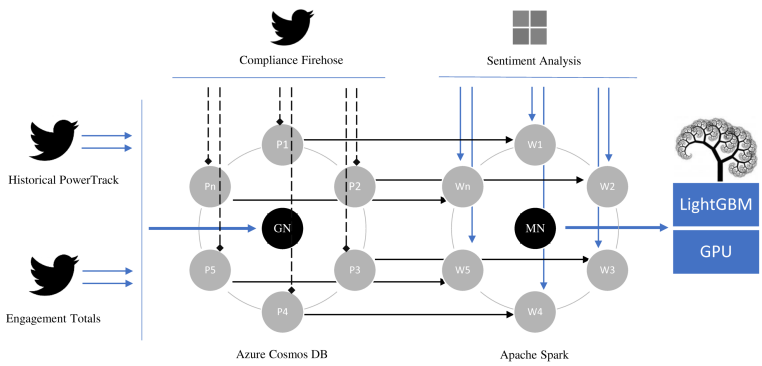


Fig. 1. Solution overview, including data acquisition, storage and analysis components. Cosmos DB gateway node GN orchestrates indexing of Twitters historical data to partitions P, for simultaneous feature extraction by Spark worker nodes W, before aggregation by master node MN for GPU accelerated predictive analysis.

3.1 Data acquisition

We use Twitters Historical APIs to acquire datasets of tweets for training and validation against other studies. In contrast to sampling Twitters x-hose, predominant in prior work, we apply Twitters PowerTrack search rules, to formulate and collect entire datasets retroactively. The documents are then stored in a globally distributed NO-SQL database, hosted by Microsoft Azure. The data remains online, exposed to every privacy request applicable.

3.2 Privacy compliant storage

Data analyzed in this study is publicly available during collection. Exactly how much of it remains public, changes rapidly afterwards. Account removal, suspension, or deleting of a single tweet render affected content unavailable for analysis in a privacy-compliant way. Users exercise their right to be forgotten at an unprecedented rate. We consume an average of 4,000 of such requests per second via Twitters Compliance Firehose API and apply to our storage simultaneously with analysis. For perspective, the average rate of new tweets published today is 8,000/s. To support this velocity and rapid feature extraction for dependent analysis we choose Azure Cosmos DB as the persistent data store.

3.3 High accuracy labels

In the contrast to prior work, we do not rely on available retweet count for training supervision. Twitter's Engagement Totals API is called during data collection, to retrieve the number of retweets and favorites ever registered for the tweet (including those deleted shortly after). This enables our data collection effort to focus on unique content only, reducing the document volume required for the task (and proportional compliance responsibility) by more than half, while ensuring 100% accuracy of the supervisory signal.

3.4 Sentiment analysis

To compute document sentiment, we adopt Text Analytics API from Microsoft Cognitive Services [27], a collection of readily consumable ML algorithms in the cloud. At the time of this study, the service supports 18 languages: English, Spanish, Portuguese, French, German, Italian, Dutch, Norwegian, Swedish, Polish, Danish, Finnish, Russian, Greek, Turkish, Arabic, Japanese and Chinese. The service is for-profit and continuously improving (changing) over time, which might challenge reproduction. To address this, we share the score of each document.

3.5 Compute

We conduct an in-memory analysis of entries no longer personally identifiable. This prevents fragmentation of sensitive data outside of the central store exposed to user privacy requests. Instead of anonymizing the datasets, sensitive

or deleted information is eliminated from storage and future analysis as soon as the request from the user is processed by the social media platform. We dedicate an Apache Spark cluster to data preprocessing and analysis. Spark is efficient at iterative computations and is thus well-suited for the development of large-scale machine learning applications [26]. Communication performance between Spark and our privacy-compliant Cosmos DB enables feature extraction at rates exceeding 65,000 tweets per second. The resulting in-memory dataset is then aggregated by the Spark master node, equipped with Tesla K80 GPUs (Graphics Processing Units) for predictive analysis and model tuning. We choose LightGBM framework to train our Gradient Boosted Regression Tree and explain the choice in the following section.

4 Data collection

We use the new framework to build multiple datasets across different time periods for training and evaluation of our models (Table 1)

Table 1. Datasets acquired

Dataset	Timeframe	Months	Language	w/images only	Total	Unique (acquired)	Never retweeted
MBI [6]	2013.02-2013.03	2	English	TRUE	2,724,764	1,319,288	1,042,411
T2015 [37]	2015.11-2016.04	6	English	TRUE	9,025,826	2,804,153	2,106,475
T2016 [37]	2016.10-2015.12	3	English	TRUE	8,469,016	2,736,600	2,088,377
T16-BIO	2015.06-2017.06	12	Multi (18x)	FALSE	27,032,417	14,788,552	12,809,021
T2017-BIO	2017.01-2018-02	14	Multi (18x)	FALSE	19,850,448	9,719,264	8,774,009

Benchmark datasets We acquire three benchmark datasets MBI, T2015 and T2016 (with a total of 6,860,041 unique tweets) to enable comparison with the work of [25,22,6,37]. The datasets match the same filters, as applied before (e.g., timeframe, language or presence of image attachment) yet result in higher volume. We follow [37,6] to split the tweets into 70% training, 10% validation, and 20% test sets respectively.

Twitter 2017 For the general multilanguage model, we have collected 10 million unique tweets and used 9.7M of them for predictive analysis, after applying privacy requests. The dataset has been downsampled from the entire Twitter 2017 volume to 18 languages supported by the sentiment scoring service, then using Twitter PowerTracks sample and bio operators, to manage the volume without sacrificing our models generalization capability over the full year.

4.1 Sentiment score and all-time totals

Retweet counts, favorite counts, and sentiment scores were collected for ca. 30 million unique tweets, simultaneously with applying privacy requests. It is worth noting that 85% of unique tweets acquired had never been retweeted.

4.2 Feature selection

Multiple features have been extracted from the rich Twitter metadata, to capture what is being said (content), by who (author), when (temporal) and how (sentiment). Table 2 describes selected features and their Pearson correlation coefficient with the logarithm of retweet count in T2017-BIO. Only the information available at the time of acquisition or immediately after is considered, to maximize the scalability of the solution. Specifically, we do not consider the early performance of the tweet (i.e., retweet or favorite counts received) or image-based features at this point.

Some authors (e.g., celebrities) receive more attention than others despite low activity. We calculate the two author ratio features in an attempt to isolate such examples. Number of attachments (like hashtags, mentions, URLs, images, symbols and videos) compete for viewers atten-tion with the original 140-character body of the tweet, and their total count is also considered. Finally, we log-transform selected author features (e.g. author's favorite and listed counts) due to power-law distribution [5].

Table 2. Feature summary

Modality	Feature	Type	Pearson
(A) Author	followersCount	ordinal	0.205920
	friendsCount	ordinal	0.082779
	accountAgeDays	ordinal	0.020379
	statusesCount	ordinal	-0.001455
	actorFavoritesCount	ordinal	0.029914
	actorListedCount	ordinal	0.221067
	actorVerified	categorical	0.202722
(C) Content	attachmentsTotal	ordinal	0.085333
	mentionCount	ordinal	-0.006590
	hashtagsCount	ordinal	0.104335
	mediaCount	ordinal	0.147623
	urlCount	ordinal	0.082549
	isQuote	categorical	0.061915
(L) Language	languageIndex	categorical	0.005199
	sentimentValue	continuous	0.059863
(T) Temporal	postedHour	ordinal	0.016639
	postedDay	ordinal	-0.000963
	postedMonth	ordinal	-0.004129
	postedDayTime	categorical	0.016639
	postedWeekDay	categorical	-0.001002

5 Methodology

We consider the problem of predicting the scale of retweet cascade for a given tweet based on data modalities available immediately after its delivery. The

author features are used together with the content, language, and temporal to predict the number of future retweets. In this study, we assume the future retweet count r of a tweet follows Poisson distribution:

$$P(R = r \mid \lambda) = \frac{e^{-\lambda}\lambda^{-r}}{r!} \tag{1}$$

where the latent variable $\lambda \in R^{+}$ defines the mean and variance of the distribution, and maximize the Poisson log-likelihood given a collection of N training tuples of tweets t_i and their retweet counts $r_{gt,i}$

$$\theta^{*} = \arg\min_{\theta} \frac{1}{N} \sum [r_{gt,i} \ln \lambda(t_i) + \lambda(t_i)] \tag{2}$$

where θ contains all parameters of the proposed model.

5.1 Gradient Boosted Regression Tree

GBRT is a tree ensemble algorithm which builds one regression tree at a time by fitting the residual of the trees that preceded it. With our twice-differentiable loss function, denoted as:

$$L_{\text{Poisson}}(r_{gt}, t) = r_{gt} \ln \lambda(t) + \lambda(t) \tag{3}$$

GBRT minimizes the loss function (regularization term omitted for simplicity):

$$L = \sum_{i=1}^{N} L_{\text{Poisson}}(r_{gt,i}, F(t_i)) \tag{4}$$

with a function estimation F(t) represented in an additive form:

$$F(t) = \sum_{m=1}^{T} f_m(t) \tag{5}$$

where each $F_m(t)$ is a regression tree and T is the number of trees. GBRT learns these regression trees in an incremental way: at m-stage, fixing the previous $m-1$ trees when learning the m-th trees. To construct the m-th tree, GBRT minimizes the following loss:

$$L_m = \sum_{t=1}^{N} L_{\text{Poisson}}(r_{gt,i}, F_{m-1}(t_i) + f_m(t_i)) \tag{6}$$

where $F_{m-1}(t) = \sum_{k}^{m-1} f_k(t)$.

The optimization problem (6) can be solved by Taylor expansion of the loss function:

$$L_m \approx \bar{L}_m = \sum_{i=0}^{N} [L_{\text{Poisson}}(r_{gt,i}, F_{m-1}(t_i)) + \nabla_i f_m(t_i) + \frac{\nabla_i^2}{2} f_m^2(t_i)] \tag{7}$$

with the gradient and Hessian defined as:

$$\begin{aligned}\nabla_i &= \frac{\partial L_{\text{Poisson}}(r_{gt,i}, F(t_i))}{\partial F(t_i)} \mid F(t_i) = F_{m-1}(t_i) \\ \nabla_i^2 &= \frac{\partial L^2_{\text{Poisson}}(r_{gt,i}, F(t_i))}{\partial^2 F(t_i)} \mid F(t_i) = F_{m-1}(t_i)\end{aligned} \tag{8}$$

We train our GBRT by minimizing $\bar{L}_m$ which is equivalent to minimizing:

$$\min_{f \in F} \sum_{i=1}^{N} \frac{\nabla_i^2}{2} (f_m(t_i) + \frac{\nabla_i}{\nabla_i^2})^2 \tag{9}$$

This approach is vulnerable to overdispersion and power-law distribution, characterizing the retweet count. In extreme cases where Hessian is nearly zero (9) approaches positive infinity. To safeguard the optimization, we cap each trees weight estimation at 1.5 and follow [5] to use total retweet count as ground-truth after log-transformation:

$$r_{gt} = \ln(r_{total} + 1) \tag{10}$$

5.2 Gradient Boosting Framework

LightGBM [21] implementation of GBDT is chosen for the task, due to distinctive techniques applicable. Experiments on multiple public datasets show that Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) can accelerate the training process by over 20 times while achieving almost the same accuracy [21]. Most of all, LightGBM implements a novel histogram-based algorithm to approximately find the best splits which is highly scalable on GPUs [40]. The framework allows us to explore substantially larger hyperparameter space during cross-validation. Finally, LightGBM offers good accuracy with integer-encoded categorical features by applying [13] to find the optimal split over categories. This often performs better than one-hot encoding and enables treating more features as categorical while avoiding dimensionality explosion.

6 Experiments

We exercise gradient boosted Poisson regression in experiments organized by datasets, to tune and compare our approach against recent state-of-the-art methods, before attempting to generalize the prediction across topics and cultures in the multilingual extended timeframe study.

6.1 Evaluation metrics

We compute the Spearman Rho ranking coefficient, to measure our models ability to rank the content by expected popularity. Interpretation of this coefficient

is domain specific, with guidelines for social/behavioral sciences proposed by [9]. SpearmanR from SciPy version 1.4.0 is used to ensure tie handling. We did not find this concern expressed in prior work. The p-value for all reported Spearman results is $p < 0.001$

Relative and absolute measures of fit: R^2, and RMSE are chosen for optimization, to penalize large error higher (i.e. when underestimating highly viral content or vice-versa). The mean-absolute-percentage-error (MAPE) is computed due to popularity in previous studies [37], but not considered for tuning. We dispute MAPEs value relative to above when fitting asymmetric, zero-inflated distribution of the dependent variable (like retweet count). It is undefined for the majority of examples (Table 1), which never receive a retweet and penalizes errors for least retweeted higher.

6.2 Validation on benchmark datasets

We begin with evaluation of our multimodal GBRT against previous state-of-the-art methods. For a fair comparison, we use Poisson regression on the joint author, content and temporal features (ACT), before including sentiment (ACTL). Table 4 demonstrates that our proposed model achieves substantially higher ranking performance, compared to other content-based methods, already before considering image and propagation modalities. Using more advanced feature representations, sentiment score and high accuracy ground-truth, we outperform the state-of-the-art by more than 37% on multiple datasets.

Table 3. Method performance on benchmark datasets.

Method	**SpearmanR**			**MAPE**		
	MBI	**T2015**	**T2016**	**MBI**	**T2015**	**T2016**
McParlene [25,37]	0.188	0.269	0.257	0.093	0.121	0.137
Khosla [22,37]	0.185	0.273	0.254	0.097	0.103	0.124
Cappallo [6,37]	0.189	0.265	0.258	0.089	0.095	0.119
Mazloom [24,37]	0.190	0.287	0.262	0.073	0.097	0.117
Wang [37]	0.229	0.358	0.350	0.057	0.084	0.103
Ours (ACT)	**0.322**	**0.498**	**0.503**	0.247	0.266	0.256
Ours (all)	**0.323**	**0.499**	**0.504**	0.247	0.266	0.255
	R^2			**RMSE**		
	MBI	**T2015**	**T2016**	**MBI**	**T2015**	**T2016**
Ours (ACT)	**0.303**	**0.417**	**0.391**	**0.444**	**0.553**	**0.555**

6.3 Multilingual, extended timeframe experiments

We apply our method to the new T2017-BIO dataset to generalize popularity prediction across languages and time. Tweet $t(A, C, T, L)$ includes content descriptions C, language descriptions L and is rst issued by author A, at the

time T. Table 4 summarizes contributions of these modalities individually and in combination. The baseline model is trained on a single feature, most popular in literature: the count of authors followers, notified about the tweet.

Table 4. Quantitative evaluation of A: actor, C: content, T: temporal, and L: language features. SpearmanR, R squared: higher is better. RMSE, MAPE: lower is better

Features	**SpearmanR**	R^2	**RMSE**	**MAPE**
A	0.310	0.317	0.359	0.133
C	**0.211**	0.055	0.422	0.160
T	0.062	0.001	0.432	0.171
L	0.164	0.017	0.430	0.167
AC	0.356	**0.396**	0.337	0.121
AT	0.311	0.316	0.359	0.132
AL	0.324	0.320	0.358	0.130
CT	0.220	0.059	0.421	0.159
CL	0.269	0.076	0.417	0.154
TL	0.170	0.019	0.430	0.166
ATL	0.324	0.320	0.358	0.130
ACT	0.357	0.395	0.338	0.120
ACL	0.369	0.399	0.336	0.119
ACTL	**0.369**	**0.402**	**0.336**	**0.118**
Baseline	**0.180**	**0.091**	**0.414**	**0.160**

7 Discussion

When prioritizing social posts by expected popularity, model's ranking performance might precede metrics of overall fit. Interpretation of Spearman and R^2 metrics is domain specific. For social/behavioral sciences, reaching 0.5 indicates strong correlation [9]. The final study aimed to explore generalizability of our method over an extended time-frame and 18 languages. The relative insignificance of the Temporal modality (Table 4) suggests low correlation between the time of posting and the content popularity, thereby challenging the common intuition, that posting at the time of audiences activity helps propagating the content. We also find that content-based features alone have higher value for expected popularity ranking than the number of followers. How many people like you appears less important than what you have to say.

Non-linear advanced ML algorithms like deep neural networks and gradient boosted decision trees are among the most successful methods used today. The fact is often attributed to the inherent capability of discovering non-linear relationships between groups of features. It was not necessary in our study to compute e.g., all cross-products to rival state-of-the-art, and at times we have noticed a higher cumulative contribution of combined modalities over their individual gains (Table 4). The size of the audience immediately exposed to the tweet, measured as the count of the authors followers, remains the single strongest predictor

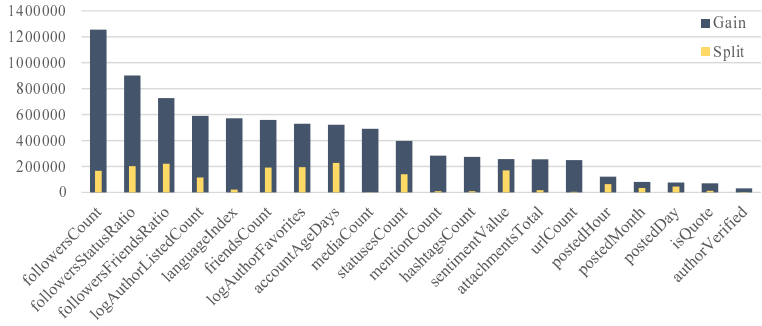


Fig. 2. Feature level importance

of tweet popularity when considered in isolation (Figure 2). The number of times an author has been listed by others, followed others or favorited other content are also among significant features, open to interpretation. Number of friends is arguably related to the diversity of content the author is exposed to. We expect the count of tweets favorited over time (i.e. age of account) to differentiate active from passive consumers. Assuming the authors influence is measured by her capacity to spread information in the social network [32], could the diversity of content actively consumed over time maximize authors influence? We propose this hypothesis for computational social science.

8 Conclusions and future work

In this paper, we have studied the problem of predicting tweet popularity under scalability, explainability and privacy compliance constraints. Our method estimates the potential reach of a tweet i.e. size of retweet cascades based on modalities available immediately after document creation. We prove it is possible to rival state-of-the-art results without compromising on explainability, scalability or privacy compliance. Our Gradient Boosted Regression Tree, combining available modalities with sentiment score and high accuracy ground-truth achieves state-of-the-art results on multiple datasets and is the first to achieve strong [9] virality ranking performance. In the final round of experiments, we apply our method to generalize prediction across extended time-frame in 18 languages and explain the contribution of each modality.

Training the final model on NVidia Tesla K80 took 10 minutes. Computing predictions for the 2 million unique tweets in the validation set, took another 45 seconds. Thats over 44,000 tweets scored per second, with a single GPU. Assuming incoming tweets are already vectorized, the ACT model deployed on Tesla K80 can cope with 5 (five) times todays Twitter volume and velocity. [37] take up to 72 additional hours (after data collection) to acquire propagation features

for the prediction. During that time, our model will have predicted popularity for up to 11 billion tweets.

8.1 Applications

Our model is ready for production with immediate application to social media monitoring. The proposed framework is extendable to other data modalities (e.g. visual) and other methods (e.g. deep neural networks) Our privacy compliant storage solution is immediately applicable to data collection and analysis from other social networks exposing privacy signal (e.g. Tumblr and WordPress, with privacy requests available as compliance interactions from DataSift). Our solution to focus analysis on temporary in-memory samples, created ad-hoc for every iteration, from a single central persistent storage to receive compliance requests, is applicable to any social network sourced data. Our solution to rely on dedicated APIs for high accuracy labels, instead of error prone counting or crawling used in prior work, is immediately applicable to Instagram, Tumblr and Facebook Pages. Our explainable GBRT approach is immediately applicable to Instagram and Tumblr.

8.2 Acknowledgements

This project is supported by Microsoft Development Center Copenhagen and the Danish Innovation Fund, Case No. 5189-00089B. We would like to thank Charlotte Mark, Lars Kai Hansen, Joerg Derungs, Petter Stengard and Uffe Kjall. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

1. Ahmed, M., Spagna, S., Huici, F.: A Peek into the Future : Predicting the Evolution of Popularity in User Generated Content. In: Proceedings of the sixth ACM international conference on Web search and data mining (2013). https://doi.org/10.1145/2433396.2433473
2. Barabsi, A.L., Psfai, M.: Network science. Cambridge University Press, Cambridge (2016), http://barabasi.com/networksciencebook/
3. Bello-Orgaz, G., Jung, J.J., Camacho, D.: Social big data: Recent achievements and new challenges. Information Fusion (2016). https://doi.org/10.1016/j.inffus.2015.08.005
4. Bunyamin, H., Tunys, T.: A Comparison of Retweet Prediction Approaches: The Superiority of Random Forest Learning Method. TELKOMNIKA (Telecommunication Computing Electronics and Control) **14**(3), 1052 (sep 2016). https://doi.org/10.12928/telkomnika.v14i3.3150, http://www.journal.uad.ac.id/index.php/TELKOMNIKA/article/view/3150
5. Can, E.F., Oktay, H., Manmatha, R.: Predicting retweet count using visual cues. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13 (2013). https://doi.org/10.1145/2505515.2507824
6. Cappallo, S., Mensink, T., Snoek, C.G.: Latent Factors of Visual Popularity Prediction. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15 (2015). https://doi.org/10.1145/2671188.2749405
7. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: ICWSM 10 (2010). https://doi.org/10.1.1.167.192
8. Cheng, J., Adamic, L.A., Dow, P.A., Kleinberg, J., Leskovec, J.: Can Cascades be Predicted? (mar 2014). https://doi.org/10.1145/2566486.2567997, http://arxiv.org/abs/1403.4608http://dx.doi.org/10.1145/2566486.2567997
9. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates (1988)
10. Dong, X., Mavroeidis, D., Calabrese, F., Frossard, P.: Multiscale event detection in social media. Data Mining and Knowledge Discovery (2015). https://doi.org/10.1007/s10618-015-0421-2
11. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (Apr 2013). https://doi.org/10.1145/2436256.2436274, http://doi.acm.org/10.1145/2436256.2436274
12. Firdaus, S.N., Ding, C., Sadeghian, A.: Retweet prediction considering user's difference as an author and retweeter. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016 (2016). https://doi.org/10.1109/ASONAM.2016.7752337
13. Fisher, W.D.: On Grouping For Maximum Homogeneity. American Statistical Association Journal (1958), http://www.csiss.org/SPACE/workshops/2004/SAC/files/fisher.pdf
14. Gan, M., Jiang, R.: FLOWER: Fusing global and local associations towards personalized social recommendation. Future Generation Computer Systems (2018). https://doi.org/10.1016/j.future.2017.02.027
15. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management (2015). https://doi.org/10.1016/j.ijinfomgt.2014.10.007

16. Hansen, L.K., Arvidsson, A., Nielsen, F.A., Colleoni, E., Etter, M.: Good friends, bad news - Affect and virality in twitter. In: Communications in Computer and Information Science (2011). https://doi.org/10.1007/978-3-642-22309-9$_5$
17. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? (dec 2017), http://arxiv.org/abs/1712.09923
18. Huang, Y., Dong, H., Yesha, Y., Zhou, S.: A Scalable System for Community Discovery in Twitter During Hurricane Sandy. In: 2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. pp. 893–899. IEEE (may 2014). https://doi.org/10.1109/CCGrid.2014.122, http://ieeexplore.ieee.org/document/6846543/
19. Ishiguro, K., Kimura, A., Takeuchi, K.: Towards automatic image understanding and mining via social curation. In: Proceedings - IEEE International Conference on Data Mining, ICDM (2012). https://doi.org/10.1109/ICDM.2012.37
20. Kaisler, S., Armour, F., Espinosa, J.A., Money, W.: Big data: Issues and challenges moving forward. In: Proceedings of the Annual Hawaii International Conference on System Sciences (2013). https://doi.org/10.1109/HICSS.2013.645
21. Ke, G., Meng, Q., Wang, T., Chen, W., Ma, W., Liu, T.Y., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems (2017). https://doi.org/10.1046/j.1365-2575.1999.00060.x
22. Khosla, A., Das Sarma, A., Hamid, R.: What makes an image popular? In: Proceedings of the 23rd international conference on World wide web - WWW '14 (2014). https://doi.org/10.1145/2566486.2567996
23. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web - WWW '10 (2010). https://doi.org/10.1145/1772690.1772751
24. Mazloom, M., Rietveld, R., Rudinac, S., Worring, M., van Dolen, W.: Multimodal Popularity Prediction of Brand-related Social Media Posts. In: Proceedings of the 2016 ACM on Multimedia Conference - MM '16 (2016). https://doi.org/10.1145/2964284.2967210
25. McParlane, P.J., Moshfeghi, Y., Jose, J.M.: "Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr. Proceedings of International Conference on Multimedia Retrieval - ICMR '14 (2014). https://doi.org/10.1145/2578726.2578776
26. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M.J., Zadeh, R., Zaharia, M., Talwalkar, A.: Mllib: Machine learning in apache spark. J. Mach. Learn. Res. **17**(1), 1235–1241 (Jan 2016), http://dl.acm.org/citation.cfm?id=2946645.2946679
27. Microsoft: Cognitive Services APIs reference. https://westus.dev.cognitive.microsoft.com/docs/services/TextAnalytics.V2.0/operations/56f30ceeeda5650db055a3c9 (2017), accessed: 2018-09-05
28. Nesi, P., Pantaleo, G., Paoli, I., Zaza, I.: Assessing the reTweet proneness of tweets: predictive models for retweeting. Multimedia Tools and Applications (2018). https://doi.org/10.1007/s11042-018-5865-0
29. Palovics, R., Daroczy, B., Benczur, A.A.: Temporal prediction of retweet count. In: 4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013 - Proceedings (2013). https://doi.org/10.1109/CogInfoCom.2013.6719254

30. Peng, H.K., Zhu, J., Piao, D., Yan, R., Zhang, Y.: Retweet modeling using conditional random fielDs. In: Proceedings - IEEE International Conference on Data Mining, ICDM (2011). https://doi.org/10.1109/ICDMW.2011.146
31. Peng, S., Zhou, Y., Cao, L., Yu, S., Niu, J., Jia, W.: Influence analysis in social networks: A survey (2018). https://doi.org/10.1016/j.jnca.2018.01.005
32. Pezzoni, F., An, J., Passarella, A., Crowcroft, J., Conti, M.: Why do I retweet it? An information propagation model for microblogs. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2013). https://doi.org/10.1007/978-3-319-03260-3$_3$1
33. Samek, W., Wiegand, T., Müller, K.R.: Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models (aug 2017), http://arxiv.org/abs/1708.08296
34. Sapountzi, A., Psannis, K.E.: Social networking data analysis tools & challenges. Future Generation Computer Systems (2018). https://doi.org/10.1016/j.future.2016.10.019
35. Sheela, L.J.: A Review of Sentiment Analysis in Twitter Data Using Hadoop. International Journal of Database Theory and Application (2016). https://doi.org/10.14257/ijdta.2016.9.1.07
36. Tan, C., Lee, L., Pang, B.: The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 175–185. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/P14-1017
37. Wang, K., Bansal, M., Frahm, J.M.: Retweet wars: Tweet popularity prediction via dynamic multimodal regression. In: Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018 (2018). https://doi.org/10.1109/WACV.2018.00204
38. Wu, B., Shen, H.: Analyzing and predicting news popularity on Twitter. International Journal of Information Management (2015). https://doi.org/10.1016/j.ijinfomgt.2015.07.003
39. Zaman, T.R., Herbrich, R., van Gael, J., Stern, D.: Predicting Information Spreading in Twitter. In: Workshop on Computational Social Science and the Wisdom of Crowds, NIPS 2010 (2010). https://doi.org/10.1016/j.jclepro.2015.12.007
40. Zhang, H., Si, S., Hsieh, C.J.: GPU-acceleration for Large-scale Tree Boosting (jun 2017), http://arxiv.org/abs/1706.08359
41. Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: SEISMIC: A self-exciting point process model for predicting tweet popularity. CoRR **abs/1506.02594** (2015), http://arxiv.org/abs/1506.02594

Appendix B

The Complexity of Social Media Response: Statistical Evidence For One-Dimensional Engagement Signal in Twitter

The Complexity of Social Media Response: Statistical Evidence For One-Dimensional Engagement Signal in Twitter

Damian Konrad Kowalczyk[1,2][a], Lars Kai Hansen[2][b]
[1]*Microsoft Development Center Copenhagen, Business Applications Group, Kanalvej 7*
[2]*Technical University of Denmark, Department of Applied Mathematics and Computer Science, Matematiktorvet 303B*
2800 Kongens Lyngby, Denmark
dakowalc@microsoft.com, {damk, lkai}@dtu.dk

Keywords: Social, Influence, Engagement, Virality, Popularity, Twitter

Abstract: Many years after online social networks exceeded our collective attention, social influence is still built on attention capital. Quality is not a prerequisite for viral spreading, yet large diffusion cascades remain the hallmark of a social influencer. Consequently, our exposure to low-quality content and questionable influence is expected to increase. Since the conception of influence maximization frameworks, multiple content performance metrics became available, albeit raising the complexity of influence analysis. In this paper, we examine and consolidate a diverse set of content engagement metrics. The correlations discovered lead us to propose a new, more holistic, one-dimensional engagement signal. We then show it is more predictable than any individual influence predictors previously investigated. Our proposed model achieves strong engagement ranking performance and is the first to explain half of the variance with features available early. We share the detailed numerical workflow to compute the new compound engagement signal. The model is immediately applicable to social media monitoring, influencer identification, campaign engagement forecasting, and curating user feeds.

1 Social media engagement

The unprecedented amount of attention aggregated by online social networks comes under intense criticism in the recent years (Bueno, 2016; Wu, 2017; Beyersdorf, 2019; Bybee and Jenkins, 2019), as billions are now exposed to low-quality content and questionable influence. Platforms like Facebook and Twitter, offer an unparalleled opportunity for influence analysis and maximization, impacting public opinion, culture, policy, and commerce (Davenport and Beck, 2001).

Extant work on influence analysis focuses on homogeneous information networks and attributes the greatest influence to authors triggering the largest diffusion cascades (Franck, 2019). When the author's influence is modeled as the ability to maximize the expected spread of information in the network (Pezzoni et al., 2013; Eshgi et al., 2019), the most desirable user-generated content is the one propagated furthest, in Twitter measured by the number of retweets. Propagation metrics however (retweet count in particular), do not capture the average individual attention received. Retweet action does not inform, e.g., if the actor has actually read the content, let alone consider the source or whether that effort was left to the followers. Meanwhile, the abundance of information to which we are exposed through online social networks is exceeding our capacity to consume it (Weng et al., 2012), let alone in a critical way. Work presented in (Weng et al., 2012; Qiu et al., 2017) shows that content quality is not a prerequisite for viral spreading, and (Lorenz-Spreen et al., 2019) shows that the competition for our attention is growing, causing individual topics to receive even shorter intervals of collective attention. Accordingly, our exposure to low-quality information and, by extension low-quality influence is increasing (Table 1). Today, the digital foot-

Table 1: Four popular tweets ranked by the most prevalent influence predictor: size of diffusion triggered in the network, in Twitter measured by the number of retweets

Tweet (body)	Retweets	Replies	Favorites
"ZOZOTOWN新春セルが史上最速で取高100を先ほ(...)"	**4.5M**	357.4K	1.3M
"HELP ME PLEASE. A MAN NEEDS HIS NUGGS"	**3.47M**	37K	0.99M
"If only Bradley's arm was longer. Best photo ever. #oscars"	**3.21M**	215K	2.29M
"No one is born hating another person because of the color of his skin or his background or his religion..."	**1.61M**	69K	4.44M

[a] https://orcid.org/0000-0002-5612-0859
[b] https://orcid.org/0000-0003-0442-5877

print of an audience goes far beyond the retweet action. Platforms like Facebook and Twitter record an increasingly diverse set of user behaviors, including number of clicks, replies or favorites (likes). Since the work of (Pezzoni et al., 2013), Twitter has made many of these metrics available to the public, inviting a more holistic approach to influence modeling, albeit rising the complexity of all dependent tasks. Consequently, few studies to date systematically investigate how to model the strength of influence in heterogeneous information networks, and the processes that drive popularity in our limited-attention world remain mostly unexplored (Franck, 2019; Weng et al., 2012).

The four Tweets in Table 1 illustrate that the mechanisms leading to high engagement are complex. In the following work, we investigate the multidimensional response of on-line audiences to understand this complexity. We examine and consolidate multiple discrete engagement metrics towards a new compound engagement signal. While the new signal is statistically motivated, we next show the relevance of the signal for understanding engagement in multiple datasets. In particular, we show that the new signal is more predictable than the individual metrics (e.g., diffusion size measured by retweet count) prevalent in literature. Our engagement model is the first to explain half of the variance with features available early, and to offer strong (Cohen, 1988) ranking performance simultaneously. We provide the workflow for calculating the new compound engagement signal from the raw count.

The contributions of this paper are summarized as follows:

1. Parallel analysis of three individual content performance signals, showing evidence of one-dimensional engagement signal on Twitter
2. new compound engagement formula, capturing over 75% of variance in available engagement signals
3. advancing feature representation of user generated content on Twitter, to consider increasingly popular 'quote tweets', validated on two real-world datasets
4. two new engagement models (response and popularity), delivering strong ranking performance
5. new state-of-the-art in virality prediction on Twitter
6. finally, a new more holistic, compound engagement model, first to explain half of the variance with content features available at the time of posting, and to offer strong ranking performance simultaneously

2 Methodology

In this section we describe the application of unsupervised learning towards contributions (1,2,6), data collection and feature extraction approach towards contribution (1,3), and the chosen supervised method towards contributions (4,5,6).

2.1 Principal Engagement Component

We acquire the multivariate set of responses forming the ground truth vector:

$$e_{gt} = \left[e_{\text{retweets}}, e_{\text{replies}}, e_{\text{favorites}}\right]^T. \tag{1}$$

Recent work on engagement modeling, e.g., (Lee et al., 2018) defines any response as a sign of engagement, effectively reducing the multivariate response to a one-dimensional signal. However, to our knowledge, the complexity of the engagement signal has not been explored more formally. While it appears credible that the population response signals,i.e., the dimensions of the of vector **e**, are highly correlated, we can test the effective dimension of the space populated by the vectors using so-called Parallel Analysis (PA) (Horn, 1965; Jorgensen and Hansen, 2011). In PA principal component analysis of the measured signals is compared with the distribution of the principal components of null data obtained by permutation under a (null) hypothesis that there is no dependency between the individual response signals. Consistent with this hypothesis, we can permute the sequence of the signals for each observation separately. In particular, we compute the upper 95%quantile for the distribution of the eigenvalues in the permuted data. Eigenvalues of the original unpermuted data set that reject the null hypothesis are considered "signal".

Principal components are computed on the response signals subject to a variance stabilization transformation,

$$e = \ln(e_{gt} + 1), \tag{2}$$

see e.g., (Can et al., 2013; Kowalczyk and Larsen, 2019).

2.2 Projection on the engagement component

Hypothesizing a one-dimensional engagement signal, we compute the value as the projection on the first principal component of the transformed data of dimension $D = 3$,

$$\mathrm{E_l} = \sum_{\mathrm{i=1}}^{\mathrm{D}} \mathrm{w_i}\left(\ln(\mathrm{e_i} + 1) - \mu_\mathrm{i}\right), \tag{3}$$

where $\mu_i = \frac{1}{N}\sum_{n=1}^{N} \hat{e}_{i,n}$ is the i'th component of the D-dimensional mean vector for a sample of size N, while w_i is the i'th component of the first principal component, computed on the same sample.

2.3 Gradient Boosted Regression Trees (GBRT)

We consider the problem of predicting audience engagement for a given tweet based on features available immediately after its delivery (Table 3). Features describing the author are used together with the content, language, and temporal descriptors to predict the size of retweet cascade, number of likes, number of replies, and the proposed compound engagement signal. GBRT is a tree ensemble algorithm that builds one regression tree at a time by fitting the residual of the trees that preceded it. The training process minimizing a chosen twice-differentiable loss function can be described as

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} L_{\mathrm{SE}}(\hat{e}_i, e_i), \tag{4}$$

where θ contains all parameters of the proposed model, N is the number of examples, and L_{SE} is the squared error of an individual prediction,

$$L_{\mathrm{SE}}(e, \hat{e}) = (e - \hat{e})^2. \tag{5}$$

We follow (Can et al., 2013; Kowalczyk and Larsen, 2019) to stabilize variance of all individual engagement signals via log-transformation as in Equation 2.

2.3.1 Gradient Boosting Framework

We use Microsoft's implementation of Gradient Boosted Decision Trees (Ke et al., 2017) for model training and tuning. LightGBM offers accurate handling of categorical features by applying (Fisher, 1958), which limits the dimensionality of our tasks.

3 Data Collection

Recent work on social network analysis reemphasizes the importance of dataset size, to make reliable predictions from representative samples. The larger the dataset, the better the accuracy and consistency of a predictive model because it minimizes the possibility of bias. However, as argued by (Agarwal et al., 2019), this intuition is incomplete. Relying solely on short timeframe samples or keyword-based crawling can produce a large dataset full of noise and irrelevant (Bhattacharya et al., 2017) data. Careful collection and filtering strategies, in addition to large-scale sampling, are critical for building datasets representative of the population and engagement modeling at scale.

3.1 Unique Tweets

We use Twitter Historical PowerTrack APIs to collect training and validation datasets described in Table 2. Retroactive filtering of Twitter archives allows close reproduction of datasets used in prior work (where still public) e.g., (Wang et al., 2018; Kowalczyk and Larsen, 2019). Historical PowerTrack API also enables near-uniform sampling across long time-frames (Figure 1), to increase the proportion of the population in a sample, as motivated by (Kim et al., 2018). Collecting a dataset similar to T2017-ML by sampling Twitter Firehose prevalent in prior work, would have taken 14 months.

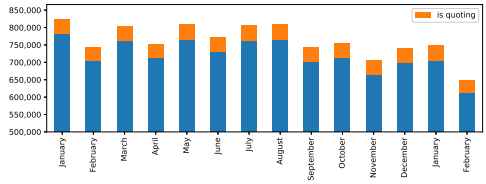


Figure 1: T2017-ML volume per month: Historical APIs allow near uniform sampling of large-scale data to ensure higher proportion of the population in a sample

3.2 Engagement totals

Three content engagement metrics are made publicly available by Twitter since 2015. We use Twitter's Engagement Totals API to retrieve the number of retweets, replies, and favorites ever registered for each tweet (even if removed later via unlike or account suspension). Use of the Engagement Totals API ensures 100% accuracy of our supervisory vector of response signals e.

3.3 Sentiment prediction

(Hansen et al., 2011; Kowalczyk and Larsen, 2019) show the impact of sentiment on tweet's virality (retweetability). We reuse sentiment predictions from (Kowalczyk and Larsen, 2019) for all tweets in the validation datasets to explore correlation with other engagement metrics and ensure fair comparison with previous results. The analysis was performed for tweets in 18 languages, using Text Analytics APIs from Microsoft Cognitive Services (Microsoft, 2017).

Table 2: Datasets acquired

Dataset	T2016-IMG	T2017-ML	T2018-ML
introduced	Wang (2018)	Kowalczyk (2018)	now
w/image only	True	False	False
languages	**English**	**18**	**all**
months total	**3**	**14**	**12**
month from	2016.10	2017.01	2018.01
unique tweets	2,848,892	9,719,264	29,883,324
quoting	421,175	583,514	2,647,072
retweets total	5,929,850	11,361,699	42,919,158
replies total	717,644	3,576,976	12,414,907
favorites total	12,665,657	29,138,707	134,523,998
no engagement	1,547,829	5,689,501	14,813,772

3.4 Datasets

Table 2 offers a summary of three datasets collected for this study.

1. **T2016-IMG** to evaluate our feature representation and training method in comparison with the work of (Mazloom et al., 2016; McParlane et al., 2014; Khosla et al., 2014; Cappallo et al., 2015; Wang et al., 2018; Kowalczyk and Larsen, 2019). The dataset matches the same filters, as applied before (timeframe, language code or the presence of an image attachment).

2. **T2017-ML** to evaluate the generalizability of our resulting models across seasons and languages (cultures) and comparison with the work of (Kowalczyk and Larsen, 2019). This dataset represents a near-uniform sample of Twitter 2017 volume in all 18 languages supported by the sentiment analysis service (Microsoft, 2017).

3. **T2018-ML** to evaluate the generalizability of our compound engagement signal across years. This dataset represents a near-uniform sample of entire Twitter 2018 volume in all known languages. In this study, T2018-ML dataset is used in unsupervised experiments only.

Datasets T2016-IMG and T2017-ML are split into 70% training, 20% test and 10% validation sets. To aid reproducibility, we share unique ID's of acquired tweets along with sentiment predictions.

3.4.1 Privacy respecting storage

The data analyzed in this study is publicly available during collection. How much of it remains public, can change rapidly afterward. We follow the architecture proposed by (Kowalczyk and Larsen, 2019) to secure the data in a central highly scalable database, exposed to applicable privacy requests from Twitter's Compliance Firehose API, and to feature extraction requests from our Spark cluster.

Table 3: Feature representation summary

Feature	Representation	Skewness	Quoted†
followers count	ordinal	0.212	**True**
friends count	ordinal	-0.321	**True**
account age (days)	ordinal	0.203	**True**
statuses count	ordinal	-0.665	**True**
actor favorites count	ordinal	-1.023	**True**
actor listed count	ordinal	0.687	**True**
actor verified	categorical	-	**True**
body length	ordinal	-1.426	**True**
mention count	ordinal	3.820	**True**
hashtag count	ordinal	5.808	**True**
media count	ordinal	3.203	**True**
url count	ordinal	1.449	**True**
language code	categorical	-	**True**
sentiment value	continuous	-0.014	False
posted hour	ordinal	-0.058	False
posted day	ordinal	0.021	False
posted month	ordinal	0.210	False
retweet count	label	6.091	n/a
reply count	label	2.330	n/a
favorite count	label	3.122	**True**

† if True, additional feature is extracted from the quoted tweet

3.4.2 Feature extraction

Table 3 describes features extracted from each tweet. To ensure scalability in production, only the information available at the time of engagement is considered. In 2015 Twitter introduced 'quote retweets' (or 'quote RTs') impacting political discourse and its diffusion as shown by (Garimella et al., 2016). Over 3.5 million tweets collected for this study quote another (Table 2). We extend the feature representation by (Kowalczyk and Larsen, 2019) to represent them. Table 3 shows in bold, an additional 14 unique features computed for quoted RT's. We log-transform highly skewed (count of followers, friends, statuses, and number of times the actor has been listed) to stabilize variance.

4 Results

We begin with examining all available content performance signals (count of retweets, replies and favorites) in the extended time-frame datasets. We look for potential correlations that could enable reducing the dimension of engagement using Parallel Analysis. In the supervised experiments, first we evaluate our methodology and feature representation against previous state-of-the-art methods, by modelling the individual influence metrics (e.g. virality) and the compound engagement on the benchmark dataset T2016-IMG. Finally we evaluate the generalizability of our method across topics and cultures, modeling engagement on the multilingual extended-timeframe dataset T2017-ML.

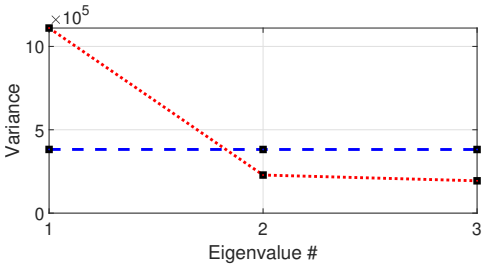


Figure 2: Parallel Analyses of the response signals for the 2017 data set provide evidence for a one-dimensional engagement signal: Only the first component ('1'- red dotted line) exceeds the 95% quantile of the corresponding eigenvalue in the null hypothesis (blue dashed line).

4.1 Evidence for a one-dimensional engagement signal

We perform Parallel Analysis and compute the principal components and their associated projected variances for the log-transformed data as well as for $Q = 100$ permutations of the data assuming the no correlation null. The one-sided upper 95% quantile is computed from the permuted samples. Variances of the un-permuted signals and the 95% quantiles for the three eigenvalues of the permuted data are shown in figure 2. Very similar results are obtained for the 2018 data set (not shown).

4.2 The engagement signal

We perform principal component analysis of the two data sets keeping a single principal component. The mean vectors and projections are found in Table 4. The variance explained by the first components in the three analyses: $2016:83\%, 2017:72\%, 2018:77\%$.

Table 4: First principal components of the extended timeframe engagement signals, used to compute the one-dimensional compound engagement (see Equation 3)

	retweets		**replies**		**favorites**	
	w_1	μ_1	w_2	μ_2	w_3	μ_3
T2017-ML	0.451	0.049	0.145	0.082	0.880	0.148
T2018-ML	0.450	0.066	0.188	0.080	0.872	0.205

4.3 Predicting Engagement

Metrics We compute the Spearman ρ ranking coefficients to measure each model's ability to rank the content depending on the definition of engagement. We compute the relative measure of fit R^2 to compare the variance explained in the compound engagement and in the individual engagement signals. The absolute measure of fit (RMSE) is chosen as an objective of optimization, to penalize large errors and relative insensitivity to outliers. The p-value for all reported ρ results is $p < 0.001$. Each metric is an average from 3-fold cross-validation. SciPy version 1.3.1 is used to ensure ρ tie handling. Interpretation of R^2 and Spearman ρ is domain-specific, with guidelines for social and behavioral sciences proposed by (Cohen, 1988).

Representation First round of our supervised experiments focus on evaluating our user-generated content feature representation and GBRT approach against previous state-of-the-art methods, in modeling established engagement signals, like the size of diffusion (e.g., retweet count), response (i.e., number of replies) and popularity (i.e., number of favorites/likes), before attempting to predict the compound engagement. Table 5 shows the performance of our GBRT with RMSE objective and new feature representation. Features extracted from the quoted content did not provide a significant boost over SOTA, likely due to visual modality dominating in the T2016-IMG dataset, as considered by (Wang et al., 2018). The approach did, however, match the performance of (Kowalczyk and Larsen, 2019) in virality ranking, and achieves strong (Cohen, 1988) performance without considering image modality. Applied to predict the new compound engagement, it sets a new benchmark for content engagement ranking $\rho = 0.680$.

Table 5: Method evaluation on the T2016-IMG dataset.

Method	R^2	ρ	RMSE
(McParlane et al., 2014)†	-	0.257	-
(Khosla et al., 2014)†	-	0.254	-
(Cappallo et al., 2015)†	-	0.258	-
(Mazloom et al., 2016)†	-	0.262	-
(Wang et al., 2018)†	-	0.350	-
(Kowalczyk and Larsen, 2019)	0.391	0.504	0.555
virality (retweets)	0.393	0.504	0.554
response (replies)	0.239	0.384	0.290
popularity (favorites)	0.500	0.656	0.665
engagement (compound)	**0.501**	**0.680**	**0.341**

† independent evaluation by (Wang et al., 2018)

Engagement The second round of supervised experiments focuses on the scalability and generalizability of our approach across topics and cultures (languages). Table 6 shows the performance of our engagement models on the multilingual extended timeframe dataset. Predicting the number of retweets with our new feature representation outperforms (Kowalczyk and Larsen, 2019), offering new state-of-the-art in virality ranking. Response and popularity models achieve strong (Cohen, 1988) ranking performance on T2017-ML. The compound engagement model again shows an increase in ranking performance over all individual engagement models, setting a new benchmark for engagement variance explained $R^2 = 0.507$.

Table 6: Engagement prediction performance on T2017-ML dataset. $SD < 0.001$ across 3-fold CV

Method	R^2	ρ	RMSE
(Kowalczyk and Larsen, 2019)	0.402	0.369	0.336
virality (retweets)	0.425	0.371	0.329
response (replies)	0.302	0.512	0.292
popularity (favorites)	0.493	0.526	0.484
engagement (compound)	**0.507**	**0.529**	**0.228**

Table 7 offers a real-world illustration of the ranking performance, in comparison with diffusion-based ranking (Table 1).

Table 7: Four popular tweets, ranked by the new compound engagement metric

Tweet (body)	**Engagement**
"No one is born hating another person because of the color of his skin or his background or his religion..."	**9.283**
"If only Bradley's arm was longer. Best photo ever. #oscars"	**9.266**
"ZOZOTOWN新春セールが史上最速で取高100を先ほ(...)"	**9.158**
"HELP ME PLEASE. A MAN NEEDS HIS NUGGS"	**8.822**

4.4 Feature Importance

Figure 3 offers a comparison of feature importance between all engagement models trained on the T2017-ML dataset. The importance equals total gains of splits which use the feature, averaged across 3-folds and rescaled to $[0,1]$ for comparison across all engagement models. The uncertainty for virality features does not exceed 6%. When predicting response (i.e., number of replies), we find the number of users mentioned to have the highest predictive value, while the number of image attachments (i.e., media count) to have almost none. The number of followers, most popular in all prior work on virality prediction is fourth when predicting compound engagement. The average number of followers received with each status or number of times the author liked another tweet is far more predictive of compound engagement.

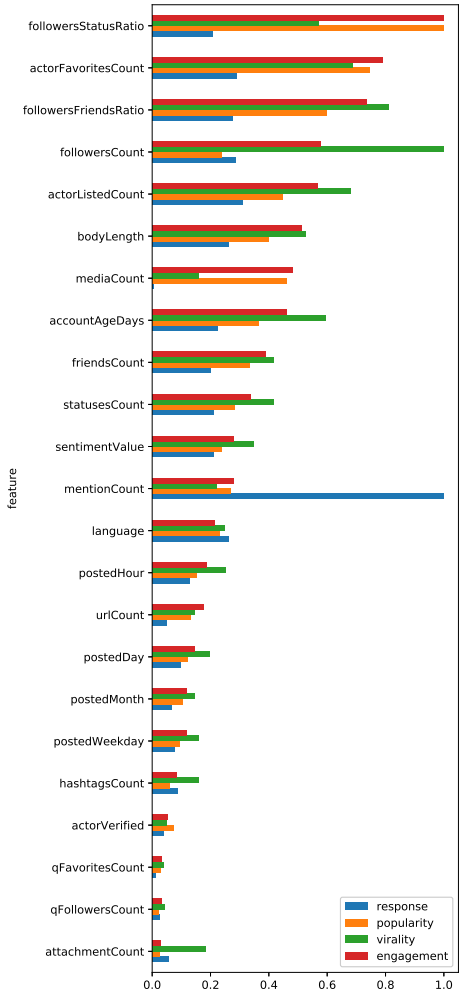


Figure 3: Relative feature importance depending on the definition of engagement (top 23 out of 31 features).

5 Conclusion

In this study, we have analyzed the complexity of the multivariate response of users engaging with social media. We have employed large-timeframe collection and filtering strategies to build datasets of unique tweets that could better represent Twitter's population. We have acquired, examined, and consolidated various response (engagement) metrics available for each of the tweets. The significant correlation found between individual response signals leads us to propose a new one-dimensional compound engagement signal. We showed on multiple benchmark datasets, that compound engagement is more predictable than any individual engagement signal, most notably the number of retweets, measuring the size of diffusion cascade, predominant in influence maximization frameworks. (Franck, 2019; Eshgi et al., 2019).

Our compound engagement model is the first to explain half of the variance with features available at the time of posting, and to offer strong (Cohen, 1988) ranking performance simultaneously. The model is ready for production with immediate application to social media monitoring, campaign engagement forecasting, influence prediction, and maximization. We

propose the ability to engage the audience as a new, more holistic baseline for social influence analysis. We share the compound engagement workflow and parameters (Eq. (3) and Table (4)) to ensure reproducibility and inspire future work on engagement modeling. We hope the future work will balance any negative impact of diffusion-based influence maximization, on our collective attention and well-being.

5.1 Acknowledgements

This project is supported by the Business Applications Group within Microsoft and the Danish Innovation Fund, Case No. 5189-00089B. We would like to acknowledge the invaluable support of Sandeep Aparajit, Jörg Derungs, Ralf Gautschi, Tomasz Janiczek, Charlotte Mark, Pushpraj Shukla, and Walter Sun. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

Agarwal, N., Dokoohaki, N., and Tokdemir, S., editors (2019). *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Lecture Notes in Social Networks. Springer International Publishing, Cham.

Beyersdorf, B. (2019). Regulating the Most Accessible Marketplace of Ideas in History: Disclosure Requirements in Online Political Advertisements after the 2016 Election. *California Law Review*, 107.

Bhattacharya, N., Arpinar, I. B., and Kursuncu, U. (2017). Real Time Evaluation of Quality of Search Terms during Query Expansion for Streaming Text Data Using Velocity and Relevance. In *Proceedings - IEEE 11th International Conference on Semantic Computing, ICSC 2017*, pages 280–281. Institute of Electrical and Electronics Engineers Inc.

Bueno, C. C. (2016). *The Attention Economy: Labour, Time and Power in Cognitive Capitalism*. Rowman & Littlefield International.

Bybee, K. J. and Jenkins, L. (2019). Free Speech, Free Press, and Fake News: What If the Marketplace of Ideas Isn't About Identifying Truth? *SSRN Electronic Journal*.

Can, E. F., Oktay, H., and Manmatha, R. (2013). Predicting retweet count using visual cues. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*.

Cappallo, S., Mensink, T., and Snoek, C. G. (2015). Latent Factors of Visual Popularity Prediction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15*.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Davenport, T. H. and Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Harvard Business Press.

Eshgi, S., Maghsudi, S., Restocchi, V., Stein, S., and Tassiulas, L. (2019). Efficient influence maximization under network uncertainty. In *INFOCOM 2019 Workshop proceedings*.

Fisher, W. D. (1958). On Grouping For Maximum Homogeneity. *American Statistical Association Journal*.

Franck, G. (2019). The economy of attention. *Journal of Sociology*, 55(1):8–19.

Garimella, K., Weber, I., and De Choudhury, M. (2016). Quote rts on twitter: Usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 200–204, New York, NY, USA. ACM.

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., and Etter, M. (2011). Good friends, bad news - Affect and virality in twitter. In *Communications in Computer and Information Science*.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Jorgensen, K. W. and Hansen, L. K. (2011). Model selection for gaussian kernel pca denoising. *IEEE transactions on neural networks and learning systems*, 23(1):163–168.

Ke, G., Meng, Q., Wang, T., Chen, W., Ma, W., Liu, T.-Y., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*.

Khosla, A., Das Sarma, A., and Hamid, R. (2014). What makes an image popular? In *Proceedings of the 23rd international conference on World wide web - WWW '14*.

Kim, H., Jang, S. M., Kim, S.-H., and Wan, A. (2018). Evaluating Sampling Methods for Content Anal-

ysis of Twitter Data. *Social Media + Society*, 4(2):205630511877283.

Kowalczyk, D. K. and Larsen, J. (2019). Scalable Privacy-Compliant Virality Prediction on Twitter. In *Proceedings of AffCon 2019 @ AAAI*, volume 2328, pages 12–27.

Lee, D., Hosanagar, K., and Nair, H. S. (2018). Advertising content and consumer engagement on social media: evidence from facebook. *Management Science*, 64(11):5105–5131.

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., and Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, 10(1):1759.

Mazloom, M., Rietveld, R., Rudinac, S., Worring, M., and van Dolen, W. (2016). Multimodal Popularity Prediction of Brand-related Social Media Posts. In *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*.

McParlane, P. J., Moshfeghi, Y., and Jose, J. M. (2014). "Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr. *Proceedings of International Conference on Multimedia Retrieval - ICMR '14*.

Microsoft (2017). Cognitive Services APIs reference. `https://westus.dev.cognitive.microsoft.com/docs/services/TextAnalytics.V2.0/`. Accessed: 2018-09-05.

Pezzoni, F., An, J., Passarella, A., Crowcroft, J., and Conti, M. (2013). Why do I retweet it? An information propagation model for microblogs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Qiu, X., F. M. Oliveira, D., Sahami Shirazi, A., Flammini, A., and Menczer, F. (2017). Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1(7):0132.

Wang, K., Bansal, M., and Frahm, J. M. (2018). Retweet wars: Tweet popularity prediction via dynamic multimodal regression. In *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*.

Weng, L., Flammini, A., Vespignani, A., and Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific reports*, 2:335.

Wu, T. (2017). *The attention merchants: The epic scramble to get inside our heads*. Vintage.

Appendix C

On the limits to multi-modal popularity prediction on Instagram: A new robust, efficient and explainable baseline

On the Limits to Multi-Modal Popularity Prediction on Instagram - A New Robust, Efficient and Explainable Baseline

Christoffer Riis*
Technical University of Denmark
Kongens Lyngby
s153147@student.dtu.dk

Damian Konrad Kowalczyk*
Business Applications Group
Microsoft Corporation
dakowalc@microsoft.com

Lars Kai Hansen
Technical University of Denmark
Kongens Lyngby
lkai@dtu.dk

ABSTRACT

The predictability of social media popularity is a topic of much scientific interest and significant practical importance. We present a new strong baseline for popularity prediction on Instagram, which is both robust and efficient to compute. The approach expands previous work by a comprehensive ablation study of the predictive power of multiple representations of the visual modality and by detailed use of explainability tools. We use transfer learning to extract visual semantics such as concepts, scenes, and objects, allowing us to interpret and explain the trained model and predictions. The study is based on one million posts extracted from Instagram. We approach the problem of popularity prediction as a ranking problem, where we predict the log-normalised number of likes. Through our ablation study design, we suggest models that outperform a previous state-of-the-art black-box method for multi-modal popularity prediction on Instagram.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Computing methodologies** → *Image representations*; • **Information systems** → Learning to rank.

KEYWORDS

popularity prediction, social media, multi-modal, explainable

ACM Reference Format:
Christoffer Riis, Damian Konrad Kowalczyk, and Lars Kai Hansen. 2020. On the Limits to Multi-Modal Popularity Prediction on Instagram - A New Robust, Efficient and Explainable Baseline. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social media platforms such as Facebook, Twitter, Flickr, and Instagram are important societal metrics. The reach of social media postings and the mechanisms determining popularity are of increasing interest for scholars of diverse disciplines. In sociology, it can be used to understand the connection between popularity and self-esteem [63]; in marketing and branding, it can clarify how to best engage and communicate with customers [14, 46, 56]; in journalism, it can be used to decide which posts to share on social media [12, 26]; and in political science, it can both be used to understand the opinion of people [30], how personalised content affect popularity [35], and what content to post to reach as many voters as possible [47]. From a data science point of view, the limits to the predictability of human behaviour is a challenging research question. In Song et al.'s seminal work on limits to mobility prediction, they argue that there is a huge gap between population and individual prediction: while individual predictability is high, population-based predictability is much harder [55]. In this paper, we focus on Instagram popularity prediction and on the hard problem of prediction using population models. Originally conceived as a photo-sharing service, visual content has been focal point of much Instagram analysis. Popularity in relation to brand value has been analysed by Mazloom et al. [41] and Mazloom et al. [40] demonstrating high predictability within specific post categories. Well-aligned with Song et al. [55], Gayberi and Oguducu [20] found very high popularity predictability of individuals' postings combining individualised models, including the given individuals' earlier postings, and an extensive multi-modal feature set. In popularity prediction, the multi-modal approaches generally give the best performance [9, 62]. However, the predictive power derived by visual information tend to lack behind other modalities [17, 24, 27].

Here, our aim is to understand the limit to predictability of Instagram popularity with population models and an eye towards both scalability and robustness. Our modelling contributions can be summarised as follows:

- Enhance the performance of the visual modality through a rich and interpretable feature set.
- Clarify the influence of different visual aspects on popularity.
- Investigate the role of four different feature sets in a comprehensive ablation study.

2 RELATED WORK

With mounting multi-modal uploads to the social media platforms, the challenge of predicting the popularity of a post suggests using different entities including metadata, author, textual, and visual information. However, the literature presents different approaches and mixed results. What is typically used is the content (also denoted *metadata*) and author information including relevant information such as e.g. the number of followers or friends a given user has. Kang et al. [29] use content and user information as input, and CatBoost [49] with data augmentation to achieve excellent

*Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

results. He et al. [24] predict popularity by adding textual information embedded with word2vec [43]. Using LightGBM [31] as their predictive model, they perform an ablation study that shows how textual content can improve performance. The ablation study also suggests that visual features extracted with ResNet-152 [23] have limited if any contribution to the performance. On the contrary, Wang et al. [62] use content, author, textual, and visual information in a joint embedding, which is then fed to a neural network and a Poisson regression model. Through an ablation study, they conclude that the content and author information indeed is the most important features, but the best performance is obtained by using all four modalities. Together, He et al. [24] and Wang et al. [62] suggest that care has to be exercised when combining modalities.

We aim to construct a new image feature extractor building upon recent work utilising deep learning [17, 24, 40, 45]. In recent years, the application of deep learning and neural networks have grown intensively as the field of computer vision has advantaged within classification [13, 50, 54, 60], object detection [7, 19, 51, 58], segmentation [2, 4, 11, 22], and generative models [18, 38, 39, 61]. Accordingly, we propose to use transfer learning with the most recent networks of computer vision to represent visual information and measure its importance in predicting popularity on social media. To improve explainability, we use embeddings formed by the input to the classifier softmax, i.e. the last layer prior to the softmax, so that each feature has a class label associated. Our work draws inspiration from earlier work using the visual information. Experiments carried out by Khosla et al. [32] find performance gains from combining low-level features (gist, texture, colour patches, gradient, and features extracted from neural networks) and semantic features such as detection of objects. Moreover, it is concluded that scenes, objects and faces are good as predictors for image popularity. McParlane et al. [42] both consider colour features, analysis of the scenery, and the number of faces in the images. Cappallo et al. [8] use visual information extracted from a pre-trained neural network, which also shows promising results for the visual modality as a descriptor for popularity prediction.

Extant recent work considers high level visual information such as concepts, scenes, and objects derived by transfer learning in the form of neural networks trained for classification or object detection tasks [20, 21, 40, 45]. Gayberi and Oguducu [20] suggest that objects and categories are important features in order to utilise the visual modality in the best way possible and therefore propose to use the MS COCO Model [6] for object detection. Gelli et al. [21] use a pre-trained network for object detection to extract high-level features and objects. Their quantitative analysis show how the visual features complement the strong information from the content and author features. Mazloom et al. [40] focus on popularity prediction within different categories such as action, animal, people, and scene. They show how human faces and animals are important for popularity prediction. Ortis et al. [45] hypothesise that semantic features of the images such as objects and scenes have an impact on the performance and therefore, they extract predictions from both Hybridnet [69] and GoogleNet [57]. Another approach is to use an image-captioning model to extract the high level information [27, 67]. Visual features include brightness, style, and colour. Quantifying the aesthetics of images in popularity prediction is seen in several papers [10, 17, 25, 41]. Chen et al. [10] propose to use Hu Moments [44] to quantify the style and colour. Ding et al. [17] use a network directly pre-trained to access the image aesthetics. Hidayati et al. [25] hypothesise that visual aesthetics are important information and, therefore, extract several high-level semantic features such as brightness, clarity, colour, and background simplicity. Mazloom et al. [41] directly extract image aesthetics as a 42-dimensional binary vector given by the content information from Instagram in the form of the feature *filter*. Another high-level feature is visual sentiment, which can be directly assessed with neural networks [21, 41]. However, we hypothesise that these features are captured in the high-level features from a deep neural network and consequently we do not apply this approach.

Table 1: Summary of the use of concepts, scenes, and objects extracted from the visual modality.

	Concepts	Scenes	Objects
Gayberi and Oguducu [20]	✓		✓
Gelli et al. [21]	✓		
Khosla et al. [32]	✓		
Mazloom et al. [40]	✓		
Mazloom et al. [41]	✓		✓
McParlane et al. [42]		✓	✓
Ortis et al. [45]	✓	✓	
Overgoor et al. [46]	✓		
Rietveld et al. [52]			✓
This study	✓	✓	✓

In multiple works, visual features are extracted implicitly by neural network embeddings pre-trained for general object recognition tasks. For example, many use a deep neural network pre-trained on ImageNet [53] for classification [17, 40, 41, 45, 46, 62, 67]. It is most common to use the embeddings from the last pooling layer with either 1024 or 2048 individual real-valued features, depending on the network structure [17, 40, 41, 46]. Ortis et al. [45] extract high-level features from three different networks by considering the last two activation layers. The three networks are pre-trained predicting classes, adjective-noun pairs, and object and scenes. Wang et al. [62] use features from a network pre-trained on ImageNet [53] and afterwards fine-tune the network for popularity prediction.

While several papers deploy transfer learning to access semantic and high-level features, recent work applies end-to-end models on the visual modality [16, 66]. Zhang and Jatowt [66] investigate the effectiveness of using neural networks in the modelling of image popularity. They hypothesise that the text features have a stronger predictive power than the visual features. With a six-layer end-to-end network, they outperform their baseline comprised of InceptionNet [59] together with Support Vector Regression and show how their network is comparable with the text-based methods word2vec [43] and GloVe [5]. Ding et al. [16] investigate the contribution of the visual content in popularity prediction by training a deep neural network to predict the intrinsic image popularity. By diving posts into different pairs giving user statistics, upload time, and captions, they train the network with a Siamese architecture. Through a qualitative analysis and a psychophysical experiment, they show how their intrinsic image popularity assessment model (IIPA) achieves human-level performance.

Our design space: Networks pre-trained for different tasks have different internal representations, which means that the high-level features will be complementary in describing images [69]. This is also seen in popularity prediction [40, 46, 52]. Therefore, we will use the deep neural network EfficientNet-B6 [60] pre-trained on ImageNet [53] for classification, Places365 ResNet-18 [68] pre-trained on the data set Places365 [68] for scene classification, and YOLOv3 [51] pre-trained on COCO [36] for object detection. We adopt the model IIPA [16] to assess the intrinsic image popularity directly. Besides introducing both EfficientNet, Places365, and YOLOv3 in popularity prediction, the novelty consists in the combination of these pre-trained models. This combination of the four complementary models gives us a strong and rich visual image representation with which we are advancing the popularity prediction on Instagram. The innovation involves the combination of specialised and state-of-the-art networks to represent both content, scenes, and objects as well as intrinsic image popularity. Additionally, these four models give us the prerequisite to understand the data and model. This strong and scalable combination puts us in a position to enhance the performance and enlarge the explainability.

There exist multiple ways to address popularity prediction on social media. Previous work predict the number of mentions for a specific event [9]; look at the popularity over time [1, 45, 64]; consider popularity for different brands [46, 52]; predict popularity for different categories [40]; define it as a binary classification problem [15, 42, 67]; but the main focus in popularity prediction on social media is to predict the number of likes, shares, views, etc., as a regression and ranking problem [10, 16, 17, 20, 24, 27, 29, 33, 34, 66]. In this paper, we will address popularity prediction as a regression and ranking problem. The different platforms give different definitions and scopes of popularity predictions. On Twitter, the focus is often on the number of retweets [33, 34, 62], but the number of likes is also used as a measurement of popularity [66]. In fact, Zhang and Jatowt [66] suggest that the number of retweets is a measure of how important a tweet is, whereas the number of likes is a measuring how sentimental a tweet is. Moreover, Kowalczyk and Hansen [34] propose a new compound signal using both the number of retweets, likes, and comments, which they show is more predictable than any individual influence predictor. On Flickr and Instagram, the literature is more consistent with the popularity signal. For research on Flickr, the number of comments [42] are used, but it is most common to use the number of views [10, 17, 21, 25, 27–29, 32, 42, 45, 64]. For research on Instagram, the number of likes is the most popular measurement of popularity [1, 3, 16, 40, 41, 46, 52, 67, 70]. However, others predict the number of comments as well [3, 52]. In this study, we will look at popularity prediction on Instagram and consequently, we will follow the majority of the literature and use the number of likes as our response variable.

3 METHODS

In this section, we first describe the 1M size data set and how it was gathered. Next, we outline the feature extraction by going through seventeen social features available at the time of posting as well as the enhanced feature extractor. Then, we define the predictive regression model in the form of LightGBM [31]. Lastly, we briefly introduce our use of the SHAP explainability tool [37].

As mentioned by several studies, there does not exist a publicly available data set for Instagram [20, 40, 46, 67]. Similar to previous studies [1, 3, 16, 20, 40, 41, 46, 52, 67, 70], we scraped Instagram and created a multi-modal data set for this study specifically. The data set consists of one million posts of type image gathered from 2018-10-31 to 2018-12-11. The size of the data set is among the largest on both Instagram and social media platforms in general, cf. Figure 1. The data set is neither categorical nor user-specific and can thus be seen as a general subset of all image posts on Instagram. However, we are aware of the inevitable bias that lies in the discard of non-public posts. The image and social information were picked up 48 hours after upload time. The author data was crawled from WWW, and afterwards used as a filter to ensure only posts from still available accounts are included in the analysis. Previous studies

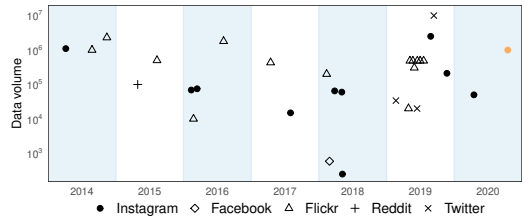


Figure 1: Different sizes of data sets have been used on the different platforms. This study (orange point) with 1 million samples is among the largest popularity prediction studies on both Instagram and social media in general. Points are shifted left or right for visual clarity.

show that the performance of popularity prediction benefits from a multi-modal approach [17, 27, 62]. Therefore, we extract features from several information sources. Overall, the features collected from each post can be divided into social features and visual features. The social features are listed in Table 2 and are branched into three categories: author, content, and temporal features. Among the author features, we extract how many followers the user has, how many other users she follows, and the number of posts the user has made. In order to stabilise the variance, we log-normalise these three variables. The transformation is given as follows by first log transforming the variable,

$$y_{log} = \log(x + 1) \tag{1}$$

and then subtracting the mean

$$y_{trans} = y_{log} - \text{mean}(y_{log}). \tag{2}$$

Furthermore, we augment the features by computing the ratios *follower per post* and *follower per following*. Regarding the content features, we extract filter (Instagram has 42 different filters), number of users tagged, whether the user liked the post, if geolocation is available, language, the number of tags, and the length of the caption measured in words and characters. From the language features, we augment the data with *is English*. Regarding the temporal features, we first extract the features consisting of the date and time

Table 2: Summary of the social features used in modelling.

Author	Skewness	Type	Origin
followers	2.728	ordinal	original
following	3.462	ordinal	original
posts	5.183	ordinal	original
follower per post	22.368	continuous	computed
follower per following	34.635	continuous	computed
Content	**Range**	**Type**	**Origin**
filter	[0, 41]	categorical	original
users tagged	[0, 20]	ordinal	original
user has liked	[0, 1]	categorical	original
has geolocation	[0, 1]	categorical	original
language	[0, 72]	categorical	original
is English	[0, 1]	categorical	computed
hashtag count	[0, 60]	ordinal	computed
word count	[0, 519]	ordinal	computed
body length	[1, 2200]	ordinal	computed
Temporal	**Range**	**Type**	**Origin**
posted day	[1, 31]	categorical	computed
posted week day	[0, 6]	categorical	computed
posted hour	[0, 23]	ordinal	computed

for posting. We omit this single feature, but do instead split it into *posted date*, *posted week day*, and *posted hour*. Lastly, since some features are not relevant for population-based popularity prediction, e.g. user id and post id are omitted. In creating a comprehensive

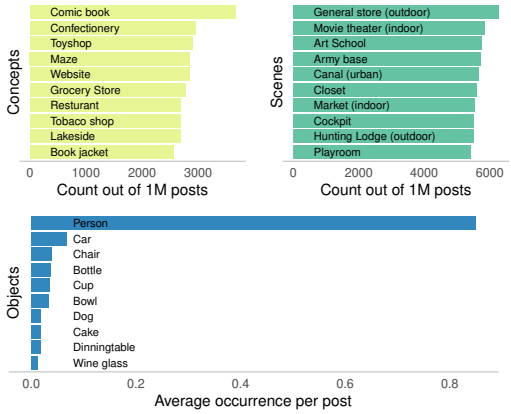


Figure 2: The ten most frequent predictions of concepts, scenes, and objects. For each category in concepts and scenes, we have counted the number of times out of the 1 million posts a category is predicted as the top-1 prediction. Since a post can have none as well as several objects, the bottom bar plot shows the average number of instances per post.

visual feature extractor, we deploy four pre-trained neural networks in order to describe concepts, scenes, objects, and intrinsic image popularity.

Concept features: To extract concept features, we use the state-of-the-art model EfficientNet-B6 [60] pre-trained on ImageNet [53]. We use the values in the last layer prior to the softmax normalization layer. This provides a 1000-dimensional vector each corresponding to a high level object class label.

Scene features: We extract a diverse set of scene features by using Places365 ResNet-18 [68]. We use the values of the last layer prior to softmax normalization. This provides a 365-dimensional interpretable vector of scene label concepts, a 102-dimensional feature vector of SUN scene attributes [48], and a single label indicating if the scene is indoors or outdoors.

Object features: YOLOv3 [51] pre-trained on COCO [36] is used to detect multiple occurences of 80 different objects. For each object, we count the number of instances providing a 80-dimensional 'bag-of-objects' histogram of object occurences.

Intrinsic image popularity: Here, we adopt the model IIPA [16] to directly assess the intrinsic image popularity in a single variable.

In total, we have 1548 features representing concepts, scenes, and objects, and one value representing the intrinsic image popularity resulting in an expressive and comprehensive visual feature representation. As part of the ablation study, we will compare these visual features without using the intrinsic image popularity (IIPA) as a predictor but instead as a baseline. To illustrate the extracted visual semantics, the top-10 concepts, scenes, and objects are seen in Figure 2. Furthermore, an example of a feature extraction is shown in Figure 3.

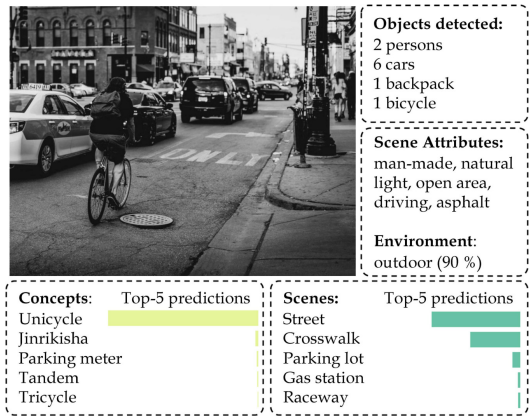


Figure 3: Example of the features extracted from an image. The associated concepts are extracted with EfficientNet, objects are detected using YOLO, and the associated scenes and scene attributes as well as the environment (indoor/outdoor) are extracted with Places365. Additionally, the image scores a neutral IIPA value at 1.96 on a normalised scale from -4 to 8, with a mean of 2.

Gradient boosting algorithms are used in social media popularity prediction [10, 20, 24, 27, 29, 33, 34] due to speed, performance and explainability. We use the framework LightGBM [31] in line with

other recent studies [24, 27, 33, 34]. LightGBM is a leaf-wise growth algorithm and uses a histogram-based algorithm to approximately find the best split. Additionally, the algorithm handles integer-encoded categorical features and uses Exclusive Feature Bundling (EFB). By combining Gradient-based One-side Sampling (GOSS) and EFB in LightGBM, Ke et al. [31] show how this algorithm can accelerate training by 20 times or more while achieving at par accuracy across multiple public data sets. The number of likes is the most popular engagement signal on Instagram [1, 3, 16, 40, 41, 46, 52, 67, 70]. We choose to predict the log-normalised number of likes (transformations from (1) and (2)) with the Spearman's Rank Correlation (SRC), Root Mean Square Error (RMSE), and R^2 as evaluation metrics.

We use *SHAP* [37] library to compute feature level explanations. Single Shapley value quantifies the effect on prediction, which is attributed to a feature. Two properties of these values make them ideal for explaining our ablation study:

Consistency and local accuracy: If we change the model s.t. a feature has a greater impact, the attribution assigned to that feature will never decrease. Features missing in the original input (i.e. removed in ablation) are attributed no importance. The values can be used to explain single predictions and to summarise the model.

Additivity of explanations: Summing the effects of all feature attributions approximates the output of the original model. Additivity, therefore, enables aggregating explanations, e.g., on a group level, towards an accurate and consistent attribution for each of the modalities in the study. We note again that our image features are conceptual (class labels). Hence, the features highlighted by SHAP can immediately be named.

We perform a basic hyper-parameter tuning of the Gradient Boosted Regression Trees offered by Ke et al. [31] on the full combination of feature groups (denoted as YIEPACT) and fix these parameters across ablation experiments to ensure fair comparison. We cap the number of leaves at 256, set the feature sampling at every iteration to 0.5 (expecting many noisy features to slow down the training otherwise), limit the number of bins when building the histograms to 255 (dictated by the GPU implementation [65]) and set the learning rate to 0.05. We train the 108 models of the ablation study (36 combinations in 3-fold CV) in a distributed environment of Apache Spark. The cluster consists of 3 nodes, each powered by a 6-core Intel Xeon CPU and an NVidia Tesla V100 GPU.

4 RESULTS AND MAIN FINDINGS

In Figure 4, the average absolute SHAP value for each feature aggregated within each group of features are displayed for each model together with the corresponding SRC. The base model CT, consisting of *Content* and *Temporal*, features achieving an SRC of 0.417 is displayed in the upper left corner. It is seen that the content features affect the prediction more than the temporal features, since the content bar is higher than the temporal bar.

Author features are essential. For the columns in Figure 4, we add author features (A), IIPA (I), and the combination of the two (IA). If we examine the first row with the base model CT, we observe that adding I to the base model increases the performance from 0.417 to 0.435 SRC, whereas adding A gives a very high increase in the performance reaching an SRC at 0.501. In fact, by looking at all rows in the second and fourth column, we see that all these models with the author features do indeed score an SRC above 0.5. The author features appear essential for reaching strong performance.

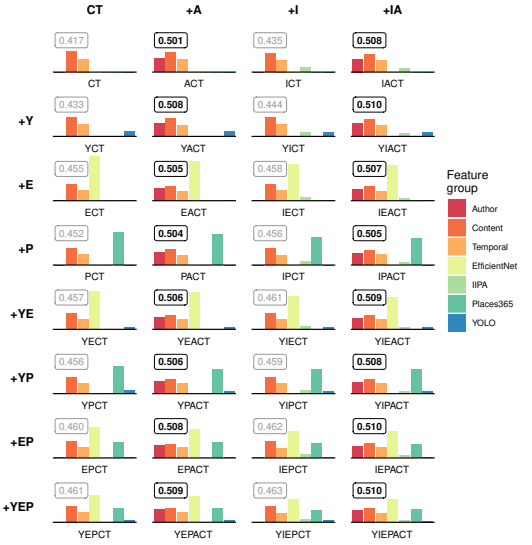


Figure 4: Average absolute SHAP value for each feature aggregated within each feature group displayed for the models. The upper left plot shows the base model with *Content* (C) and *Temporal* (T) features. In the three columns, *Author* (A) and *IIPA* (I) features are added, and for each row the groups *EfficientNet* (E), *Places365* (P), and *YOLO* (Y) - corresponding to concepts, scenes, and objects respectively - are added. For each model, the Spearman's Rank Correlation is shown.

EfficientNet has the largest effect on the predictions. In the rows below, the base model CT in Figure 4, the different semantic concepts, scenes, and objects are added to the model in the form of *EfficientNet* (E), *Places365* (P), and *YOLO* (Y) resp. Comparing the three models YCT, ECT, and PCT, it is seen that E on average, has the largest effect on the predictions. In the lower half of the column, we have the models combining these features, and again it appears that E has the largest effect. However, it should be noted that E has 1000 features, whereas P and Y only have 468 and 80 resp. In other words, the features in E combined affects the prediction more, but a single feature from P and Y might contribute more than a single feature from E. If we examine the other columns, it is indeed observed that EfficientNet on average has the largest effect on the predictions across all models.

Visual semantics are correlated. Adding combinations of the semantic groups gives a decrease in the contribution for a single group, e.g. in YEPCT the effect of both E, P, and Y are lower than for

the other models in this column in Figure 4. At the same time, we see that the SRC is increased every time new features are added to the model, indicating that the different features are complementary. However, the decrease in the different bars together with the increase in the SRC also indicate that the groups are slightly correlated and that the model might learn a better representation such that some of the features within the different groups are disregarded. In other words, this illustrates the synergy between the groups and how some features are substituted by including other features. These observations can be validated across the other columns.

If we briefly examine the three other columns, the second row shows that the effect of Y decreases as we add A, I, and IA. The same is true for E and P as the model is combining the visual semantics. In fact, the more features we combine, the lower is the contribution from each feature group. In particular, the largest model YIEPACT has the lowest contributions for each feature group.

Object detection works better with author features. In the second column in Figure 4, we add the author features (A) to the base model CT. We observe a sudden increase in the performance reaching an SRC at 0.501, which is 0.04 higher than the best model YEPCT from the first column. We have already conducted this performance increase obtained by adding author features and will, therefore, instead examine the effect of A on the visual semantics here. But first, note that models with EfficientNet features (E) always give the same or better performance than Places365 features (P) across all models, e.g. YI**E**ACT has a higher SRC than YI**P**ACT. If we examine the models without A starting with the first column, we see that the increase in performance is higher when adding E or P instead of Y, e.g. the model EPCT achieve a higher SRC than both YECT and YPCT. The same patterns are seen in the third column. However, if we examine the models with A starting with the second column, the pattern is more cluttered, since YACT achieves a higher SRC than both EACT and PACT. Moreover, we see that adding either E or P to YACT results in a decrease in performance, but adding all them in YEPACT gives the highest performance in this column. Furthermore, we do see that the combination of EP in EPACT achieves the same performance as YACT.

From these observations, we hypothesise that if you should only add one semantic feature to ACT, then Y gives the highest performance, but if you can add two features, then E and P will be equally good. However, there is no performance gain in using EP instead of Y. Lastly, even though both YEACT and YPACT have lower performance than YACT, adding all three visual semantics in YEPACT gives a small increase in performance. These hypotheses are validated by the fourth column, where again Y as a single feature is better than E and P, but adding the combination EP gives similar performance to adding Y. However, here, no performance gain is obtained by combing YIACT and IEPACT into YIEPACT. All these three models achieve the highest observed SRC at 0.510. In summary, we see how objects together with authors features are very powerful, but also how the combination of concepts and scenes is indeed powerful with and without author features.

In the following, we will investigate the features affecting the prediction most by finding the top-30 most prominent features based on the average absolute SHAP value across all models. More precisely, we aggregate the average absolute SHAP value for each feature across all models, and then divide by the number of times that feature is present in the models.

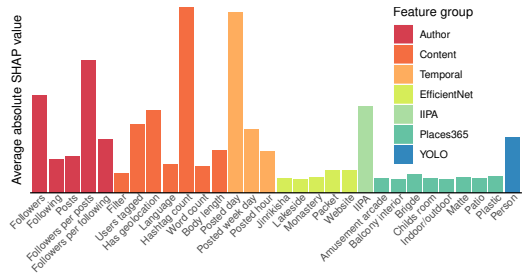


Figure 5: Average absolute SHAP value for top 30 features. The features are chosen by highest average absolute SHAP values across all models, but normalised by the number of appearances of that feature group.

In Figure 5 the top-30 features are shown coloured after each feature group. The two features *hashtag count* and *posted day* by far have the largest average absolute SHAP value and thereby affect a prediction the most. The author features *followers* and *followers per post* come right after with high contribution as well. Note how the two computed ratios *followers per post* and *followers per following* both are high and are actually affecting the prediction more than the two features *following* and *posts*. The three temporal features all have a high effect on the prediction which both shows that the day of the week and the time of the day is important information for predicting the popularity. The content features *users tagged* and *has geolocation* also have a relatively high effect.

Among the visual features, IIPA and *Person* have the largest effect and are both comparable to the social features. Yet, in general, all the visual features have a smaller effect than the social features. The social features are explained using the SHAP values individually. We summarise the SHAP values in two numbers computed as the mean of all positive and all negative SHAP values separately. In this way, we both preserve the sign and the deviation of the SHAP values. In contrast, SHAP values of different signs will cancel out each other in a regular mean calculation. In Figure 6, the positive and negative mean SHAP values for the social features are visualised.

Hashtag count and posted day are good discriminators. In Figure 6, the base model CT consisting of content and temporal features indicate that *hashtag count* and *posted day* are good discriminators. The reason is two-fold: firstly, they have high positive and negative means (e.g. the bars are large) and secondly, the magnitude of the positive and negative mean is similar, meaning that features can affect a prediction in a positive and negative direction, equally. The feature *users tagged* also has a high impact on the prediction, but the effect is mainly in a positive direction, since the positive mean is of larger magnitude than the negative mean and, consequently, it is not as good a discriminator as the two aforementioned. Moreover, *has geolocation* seems to be a good discriminator, *filter* mainly affects the prediction in a negative direction, whereas

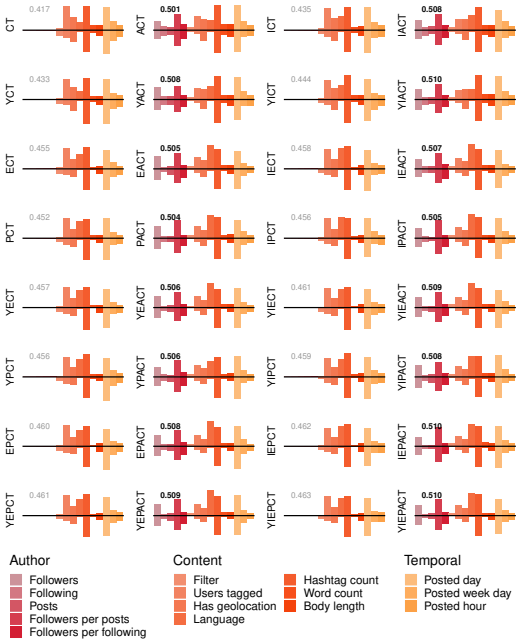


Figure 6: Average positive and negative SHAP values for most prominent social features displayed for each model.

language mainly affects the prediction in a positive direction. Regarding the size of the bars, similar trends from the top features in Figure 5 are observed in the figure.

Language is important with visual semantics. If we consider the first column in Figure 6, only small changes are observed down the rows. The size of the bars is decreasing slightly as we add visual features, e.g *word count* is larger in CT than YEPCT. Adding objects (Y) only seem to have very small effects on the bars and is not changing the relative distribution, whereas adding concepts (E) and scenes (P) give an increase in the positive mean of *language*. In fact, all the features are smaller in YEPCT than in CT except *language*, which is slightly higher. This indicates that *language* is more important when visual semantics are added to the model. The other columns validate the observation. We hypothesise that the visual predictors of popularity vary across cultures.

The caption is less important with visual features. If we compare the models in the first row with the models in the last row in Figure 6, attribution of the feature *word count* has decreased. This indicates a connection between the visual features and the word count, which suggests that the visual information can partly substitute the information in the word count. Word count is the number of words in the caption, and thus, we observe how the caption is less important when visual features are present.

Language is more important with IIPA. In the last two columns in Figure 6, IIPA is added to CT and ACT, resp. Like E and P, IIPA also affects the positive mean of *language* in a positive direction, e.g. comparing CT with ICT. This is also seen for other rows though the increase is smaller due to the increase from E and P. Therefore, we observe that language is more important with IIPA, suggesting that the definition of intrinsic popularity varies across cultures.

Visual features have a small impact on social features. Overall, only small changes are observed across the models in Figure 6, indicating that the visual features only have a small effect on the impact of social features on a prediction. If we compare the models in the first row with the models in the last row, the features *language* has increased and *word count* has decreased. If we compare ACT with YIEPACT, it is observed that the majority of the features have a smaller impact and *word count* is very small but the two author features *followers* and *followers per post* are unchanged, and the content feature *language* is actually larger. This suggests that author features are important no matter the visual information, that *language* might capture some sort of user segment, and that *word count* and visual information are highly related.

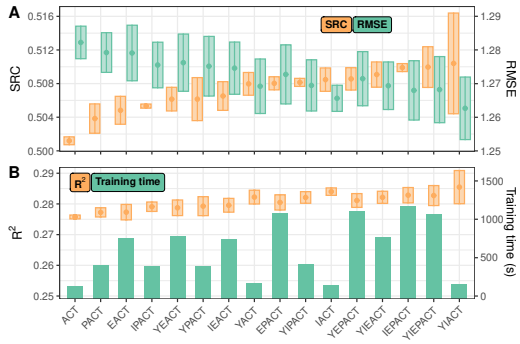


Figure 7: Performance for models getting an SRC higher than 0.5. The boxes shows ±2 standard deviations. (A) Spearman's Rank Correlation (SRC) and Root Mean Square Error (RMSE). (B) R^2 and training time.

The performance of the models is quantified using Spearman's rank correlation (SRC), Root Mean Square Error (RMSE), the R^2, and the training time. In the top panel of Figure 7, the performance ±2 standard deviations for the 16 best models are shown. As expected, the SRC and RMSE are inversely related. The standard deviations of performance between crossvalidation folds form a conservative (too large) estimate of the standard error of the mean. YIACT has the highest SRC, but also a high standard deviation, while the model IEPACT has similar performance but more robust. If we also include the R^2 and the training time in the bottom panel of Figure 7, we note that the models ACT, YACT, IACT, and YIACT are fast with training times below 200 seconds. All the other models have more than four times as many features, which is reflected in the increased

training time. If R^2 is also taken into account, YIACT has the highest values but IACT has similar performance with much lower standard deviation. The model IACT enjoys a low training time, a high R^2, and a high SRC with a small confidence interval. Hence, it is a good candidate for a strong, robust, and efficient baseline for Instagram popularity prediction. If we accept the somewhat larger training time (about 20 minutes), the model IEPACT is an excellent and robust candidate with a strong, consistent SRC performance across cross-validation folds. For a real-time application, the prediction time is a central metric. The prediction time includes the feature extraction, and we assume that if you want to predict the popularity of a new post, you have the image, content and temporal information at hand. The author features are crawled from WWW and the visual features are obtained via a propagation through the networks. In parallel, all LightGBM models run in less than one tenth of a millisecond. In Table 3 and Table 4, the prediction time for a single evaluation of a post is seen. Though the author features are relatively slow, they are vital for good performance. The visual features contribute only slightly to the prediction time and therefore, they do not change the conclusion.

Table 3: Ablation study with features groups removed. Performance metrics are given by Spearman's rank correlation (SRC) and root mean square error (RMSE) together with the training and prediction time. All standard deviations with respect to RSME and SRC are below 0.002.

	Perf.		Time	
Group removed	SRC	RMSE	Train (s)	Pred. (ms)
Author	0.463	1.202	1075	186
EfficientNet	0.509	1.158	421	1055
Places365	0.509	1.158	772	1111
YOLOv3	0.510	1.157	1170	1051
IIPA	0.509	1.159	1105	1104

5 CONCLUSION

In this paper, we address the hard problem of multi-modal popularity prediction in Instagram using population wide models. We design a comprehensive ablation study including transfer learning to represent visual semantics with the explainable features concepts, scenes, and objects. The approach is strong, since we show robustness and consistency across models that take advantage of the synergy between the visual semantics. Additionally, the lower bounds of the Spearman's rank correlation above 0.5 on a generalisable data set without the usage of a given user's earlier popularity. The approach is explainable both on a high-level based on feature groups and on low-level with individual features. We use SHAP analysis to quantify the feature importance. In particular, we find that object detection works better with author features, and language is important with visual semantics. Based on the many combinations of multi-modal models, we can make these recommendations: If training time is of importance, we recommend a model (IACT) that combines author, content and temporal features with a single dimension measure of image popularity. This model trains in less than three minutes. If the focus is on robust performance and less on time to train, we recommend a model (IEPACT) that combines author, content, temporal, intrinsic image popularity, with the EfficientNet and Places visual embeddings, which is about seven times slower in training. However, the latter model enjoys both high and consistent Spearman's rank correlation across cross-validation folds.

Table 4: Quantitative evaluation of all models given by Spearman's rank correlation (SRC), root mean square error (RMSE), R squared (R^2), and the prediction time given in milliseconds. Abbreviations: author (A), content (C), temporal (T), EfficientNet (E), Places365 (P), YOLOv3 (Y), and IIPA (I).

	SRC		RMSE		R^2		Time
Features	μ	σ	μ	σ	μ	σ	Pred. (ms)
T	0.261	0.001	1.306	0.001	0.086	0.001	<1
C	0.305	0.002	1.291	0.001	0.108	0.001	<1
A	0.349	0.002	1.266	0.001	0.141	0.001	935
CT	0.417	0.001	1.231	0.001	0.188	0.000	<1
AT	0.425	0.001	1.219	0.002	0.204	0.001	936
AC	0.426	0.000	1.216	0.001	0.207	0.000	936
CT							
YCT	0.433	0.000	1.222	0.001	0.200	0.000	71
ICT	0.435	0.001	1.219	0.001	0.204	0.000	18
YICT	0.444	0.001	1.214	0.001	0.211	0.001	88
PCT	0.452	0.001	1.210	0.001	0.216	0.001	33
ECT	0.455	0.000	1.208	0.001	0.219	0.001	89
YPCT	0.456	0.000	1.207	0.002	0.220	0.001	103
IPCT	0.456	0.000	1.206	0.001	0.221	0.001	50
YECT	0.457	0.000	1.206	0.002	0.221	0.001	159
IECT	0.458	0.001	1.205	0.001	0.222	0.000	106
YIPCT	0.459	0.000	1.204	0.001	0.224	0.001	120
EPCT	0.460	0.001	1.205	0.001	0.223	0.000	99
YIECT	0.461	0.000	1.204	0.001	0.224	0.001	176
YEPCT	0.461	0.000	1.204	0.002	0.224	0.001	169
IEPCT	0.462	0.001	1.202	0.001	0.226	0.001	116
YIEPCT	0.463	0.000	1.202	0.001	0.227	0.001	186
ACT							
ACT	0.501	0.000	1.163	0.001	0.276	0.000	936
PACT	0.504	0.001	1.162	0.001	0.277	0.001	968
EACT	0.505	0.001	1.162	0.002	0.277	0.001	1024
IPACT	0.505	0.000	1.160	0.001	0.279	0.001	985
YEACT	0.506	0.001	1.160	0.002	0.279	0.001	1094
YPACT	0.506	0.001	1.160	0.002	0.279	0.002	1038
IEACT	0.507	0.001	1.160	0.002	0.280	0.001	1041
YACT	0.508	0.001	1.158	0.002	0.282	0.001	1006
EPACT	0.508	0.000	1.159	0.002	0.280	0.001	1034
YIPACT	0.508	0.000	1.158	0.002	0.282	0.001	1055
IACT	0.508	0.001	1.156	0.001	0.284	0.001	954
YEPACT	0.509	0.001	1.159	0.002	0.281	0.001	1104
YIEACT	0.509	0.001	1.158	0.001	0.282	0.001	1111
IEPACT	**0.510**	0.000	1.157	0.002	0.283	0.001	1051
YIEPACT	**0.510**	0.001	1.157	0.002	0.283	0.002	1121
YIACT	**0.510**	0.003	**1.155**	0.002	**0.285**	0.003	1023

ACKNOWLEDGMENTS

This project is supported by the Business Applications Group within Microsoft and the Danish Innovation Fund, Case No. 5189-00089B.

REFERENCES

[1] Khaled Almgren, Jeongkyu Lee, and Minkyu Kim. 2016. Prediction of image popularity over time on social media networks. *2016 Annual Connecticut Conference on Industrial Electronics, Technology and Automation, CT-IETA 2016* (2016), 1–6. https://doi.org/10.1109/CT-IETA.2016.7868253

[2] V Badrinarayanan, A Kendall IEEE transactions on ..., and Undefined 2017. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *ieeexplore.ieee.org* (2017). https://ieeexplore.ieee.org/abstract/document/7803544/

[3] Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. *Conference on Human Factors in Computing Systems - Proceedings* (2014), 965–974. https://doi.org/10.1145/2556288.2557403

[4] Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. 2017. Unstructured Point Cloud Semantic Labeling Using Deep Segmentation Networks. *3DOR* 2 (2017), 7.

[5] Paul M. Brennan, James J.M. Loan, Neil Watson, Pragnesh M. Bhatt, and Peter Alwyn Bodkin. 2017. Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery. *British Journal of Neurosurgery* 31, 6 (2017), 682–687. https://doi.org/10.1080/02688697.2017.1354122

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. COCO-Stuff: Thing and Stuff Classes in Context. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), 1209–1218. https://doi.org/10.1109/CVPR.2018.00132 arXiv:1612.03716

[7] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. 2016. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*. Springer, 354–370.

[8] Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. 2015. Latent factors of visual popularity prediction. *ICMR 2015 - Proceedings of the 2015 ACM International Conference on Multimedia Retrieval* (2015), 195–202. https://doi.org/10.1145/2671188.2749405

[9] Guandan Chen, Qingchao Kong, Nan Xu, and Wenji Mao. 2019. NPP: A neural popularity prediction model for social media content. *Neurocomputing* 333 (2019), 221–230. https://doi.org/10.1016/j.neucom.2018.12.039

[10] Junhong Chen, Dayong Liang, Zhanmo Zhu, Xiaojing Zhou, Zihan Ye, and Xiuyun Mo. 2019. Social media popularity prediction based on visual-textual features with XGboost. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (2019), 2692–2696. https://doi.org/10.1145/3343031.3356072

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).

[12] Aman Chopra, Ashray Dimri, and Sumanu Rawat. 2019. Comparative Analysis of Statistical Classifiers for Predicting News Popularity on Social Web. *2019 International Conference on Computer Communication and Informatics, ICCCI 2019* (2019), 1–8. https://doi.org/10.1109/ICCCI.2019.8822230

[13] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3642–3649.

[14] Lisette De Vries, Sonja Gensler, and Peter S.H. Leeflang. 2012. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *Journal of Interactive Marketing* 26, 2 (2012), 83–91. https://doi.org/10.1016/j.intmar.2012.01.003

[15] Arturo Deza and Devi Parikh. 2015. Understanding image virality. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June (2015), 1818–1826. https://doi.org/10.1109/CVPR.2015.7298791 arXiv:1503.02318

[16] Keyan Ding, Kede Ma, and Shiqi Wang. 2019. Intrinsic image popularity assessment. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* October (2019), 1979–1987. https://doi.org/10.1145/3343031.3351007 arXiv:1907.01985

[17] Keyan Ding, Ronggang Wang, and Shiqi Wang. 2019. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (2019), 2682–2686. https://doi.org/10.1145/3343031.3356062

[18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2019. Density estimation using real NVP. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (2019). arXiv:1605.08803

[19] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. 2014. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2147–2154.

[20] Mehmetcan Gayberi and Sule Gunduz Oguducu. 2019. Popularity prediction of posts in social networks based on user, post and image features. *11th International Conference on Management of Digital EcoSystems, MEDES 2019* (2019), 9–15. https://doi.org/10.1145/3297662.3365812

[21] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference* (2015), 907–910. https://doi.org/10.1145/2733373.2806361

[22] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. 2017. Brain tumor segmentation with deep neural networks. *Medical image analysis* 35 (2017), 18–31.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90 arXiv:1512.03385

[24] Ziliang He, Zijian He, Jiahong Wu, and Zhenguo Yang. 2019. Feature construction for posts and users combined with lightgBM for social media popularity prediction. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (2019), 2672–2676. https://doi.org/10.1145/3343031.3356054

[25] Shintami Chusnul Hidayati, Yi Ling Chen, Chao Lung Yang, and Kai Lung Hua. 2017. Popularity meter: An influence- and aesthetics-aware social media popularity predictor. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference* (2017), 1918–1923. https://doi.org/10.1145/3123266.3127903

[26] Md Arafat Hossain, Didar Hossain Sagar, Kawsar Ahmed, Bikash Kumar Paul, Md Zamilur Rahman, and Md Ahsan Habib. 2019. Popularity prediction of online news item based on social media response. *2019 Joint 8th International Conference on Informatics, Electronics and Vision, ICIEV 2019 and 3rd International Conference on Imaging, Vision and Pattern Recognition, icIVPR 2019 with International Conference on Activity and Behavior Computing, ABC 2019* (2019), 173–177. https://doi.org/10.1109/ICIEV.2019.8858525

[27] Chih Chung Hsu, Jun Yi Lee, Li Wei Kang, Zhong Xuan Zhang, Chia Yen Lee, and Shao Min Wu. 2019. Popularity prediction of social media based on multi-modal feature mining. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (2019), 2687–2691. https://doi.org/10.1145/3343031.3356064

[28] Jiani Hu, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016. Multimodal learning for image popularity prediction on social media. *2016 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2016* (2016), 1–2. https://doi.org/10.1109/ICCE-TW.2016.7521017 arXiv:arXiv:1503.01817

[29] Peipei Kang, Zehang Lin, Shaohua Teng, Guipeng Zhang, Lingni Guo, and Wei Zhang. 2019. Catboost-based framework with additional user information for social media popularity prediction. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (2019), 2677–2681. https://doi.org/10.1145/3343031.3356060

[30] Amir Karami and Aida Elkouri. 2019. Political Popularity Analysis in Social Media. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11420 LNCS (2019), 456–465. https://doi.org/10.1007/978-3-030-15742-5_44 arXiv:1812.03258

[31] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 2017-Decem, Nips (2017), 3147–3155.

[32] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular? (2014), 867–876. https://doi.org/10.1145/2566486.2567996

[33] Damian Konrad Kowalczyk. 2019. *Scalable Privacy-Compliant Virality Prediction on Twitter*. Technical Report.

[34] Damian Konrad Kowalczyk and Lars Kai Hansen. 2019. The Complexity of Social Media Response: Statistical Evidence For One-Dimensional Engagement Signal in Twitter. (2019), 2013–2019. arXiv:1910.02807 http://arxiv.org/abs/1910.02807

[35] Anders Olof Larsson. 2019. Skiing all the way to the polls: Exploring the popularity of personalized posts on political Instagram accounts. *Convergence* 25, 5-6 (2019), 1096–1110. https://doi.org/10.1177/1354856517741132

[36] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS, PART 5 (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48 arXiv:1405.0312

[37] Scott M. Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 2017-December, Section 2 (2017), 4766–4775. arXiv:1705.07874

[38] Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. 2019. BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling. NeurIPS (2019). arXiv:1902.02102 http://arxiv.org/abs/1902.02102

[39] Shahin Mahdizadehaghdam, Ashkan Panahi, and Hamid Krim. 2019. Sparse generative adversarial network. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019* (2019), 3063–3071. https://doi.org/10.1109/ICCVW.2019.00369 arXiv:1908.08930

[40] Masoud Mazloom, Iliana Pappi, and Marcel Worring. 2018. Category Specific Post Popularity Prediction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10704 LNCS (2018), 594–607. https://doi.org/10.1007/978-3-319-73603-7_48

[41] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. 2016. Multimodal popularity prediction of brand-related social media posts. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*

(2016), 197–201. https://doi.org/10.1145/2964284.2967210

[42] Philip J. McParlane, Yashar Moshfeghi, and Joemon M. Jose. 2014. "Nobody comes here anymore, it's too crowded"; predicting image popularity on Flickr. *ICMR 2014 - Proceedings of the ACM International Conference on Multimedia Retrieval 2014* (2014), 385–391. https://doi.org/10.1145/2578726.2578776

[43] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* (2013), 1–12. arXiv:1301.3781

[44] Ming-Kuei Hu. 1962. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory* 8, 2 (1962), 179–187.

[45] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. 2019. Predicting Social Image Popularity Dynamics at Time Zero. *IEEE Access* 7 (2019), 171691–171706. https://doi.org/10.1109/ACCESS.2019.2953856

[46] Gijs Overgoor, Masoud Mazloom, Marcel Worring, Robert Rietveld, and Willemijn Van Dolen. 2017. A spatio-temporal category representation for brand popularity prediction. *ICMR 2017 - Proceedings of the 2017 ACM International Conference on Multimedia Retrieval* (2017), 233–241. https://doi.org/10.1145/3078971.3078998

[47] Ethan Pancer and Maxwell Poole. 2016. The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 U.S. presidential nominees' tweets. *Social Influence* 11, 4 (2016), 259–270. https://doi.org/10.1080/15534510.2016.1265582

[48] Genevieve Patterson and James Hays. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2012), 2751–2758. https://doi.org/10.1109/CVPR.2012.6247998

[49] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems* 2018-December, Section 4 (2018), 6638–6648. arXiv:1706.09516

[50] Waseem Rawat and Zenghui Wang. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* 29, 9 (2017), 2352–2449.

[51] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. (2018). arXiv:1804.02767 http://arxiv.org/abs/1804.02767

[52] Robert Rietveld, Willemijn van Dolen, Masoud Mazloom, and Marcel Worring. 2020. What You Feel, Is What You Like Influence of Message Appeals on Customer Engagement on Instagram. *Journal of Interactive Marketing* 49 (2020), 20–53. https://doi.org/10.1016/j.intmar.2019.06.003

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y arXiv:1409.0575

[54] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[55] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[56] Kunal Swani, George R. Milne, Brian P. Brown, A. George Assaf, and Naveen Donthu. 2017. What messages to post? Evaluating the popularity of social media communications in business versus consumer markets. *Industrial Marketing Management* 62 (2017), 77–87. https://doi.org/10.1016/j.indmarman.2016.07.006

[57] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June-2015 (2015), 1–9. https://doi.org/10.1109/CVPR.2015.7298594 arXiv:1409.4842

[58] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. 2013. Deep neural networks for object detection. In *Advances in neural information processing systems*. 2553–2561.

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December (2016), 2818–2826. https://doi.org/10.1109/CVPR.2016.308 arXiv:1512.00567

[60] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2019). arXiv:1905.11946 http://arxiv.org/abs/1905.11946

[61] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. *33rd International Conference on Machine Learning, ICML 2016* 4 (2016), 2611–2620. https://doi.org/10.4249/scholarpedia.1888 arXiv:1601.06759

[62] Ke Wang, Mohit Bansal, and Jan Michael Frahm. 2018. Retweet wars: Tweet popularity prediction via dynamic multimodal regression. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018* 2018-Janua (2018), 1842–1851. https://doi.org/10.1109/WACV.2018.00204

[63] Ruoxu Wang, Fan Yang, and Michel M. Haigh. 2017. Let me take a selfie: Exploring the psychological effects of posting and viewing selfies and groupies on social media. *Telematics and Informatics* 34, 4 (2017), 274–283. https://doi.org/10.1016/j.tele.2016.07.004

[64] Bo Wu, Wen Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time matters: Multi-scale temporalization of social media popularity. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference* (2016), 1336–1344. https://doi.org/10.1145/2964284.2964335 arXiv:1801.05853

[65] Huan Zhang, Si Si, and Cho-Jui Hsieh. 2017. GPU-acceleration for Large-scale Tree Boosting. (jun 2017). arXiv:1706.08359 http://arxiv.org/abs/1706.08359

[66] Yihong Zhang and Adam Jatowt. 2019. Image tweet popularity prediction with convolutional neural network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11437 LNCS. Springer Verlag, 803–809. https://doi.org/10.1007/978-3-030-15712-8_56

[67] Zhongping Zhang, Tianlang Chen, Zheng Zhou, Jiaxin Li, and Jiebo Luo. 2018. How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018* (2018), 2383–2392. https://doi.org/10.1109/BigData.2018.8622461 arXiv:1809.09314

[68] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

[69] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database - Supplementary Materials. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems* 1 (2014), 487–495.

[70] Alireza Zohourian, Hedieh Sajedi, and Arefeh Yavary. 2018. Popularity prediction of images and videos on Instagram. *2018 4th International Conference on Web Research, ICWR 2018* (2018), 111–117. https://doi.org/10.1109/ICWR.2018.8387246

Appendix D

Commercial Potential

(as seen in 2016)

The age of the customer is here. More and more companies realize that their competitive position depends on customer insight and experience [Sor15], due to their strong correlation with loyalty and revenue. According to a recent survey, companies around the world are increasingly looking to Big Data analytics to help them boost consumer and end-user satisfaction through positive experience in every phase of sales, service and support [BB15, BSB14]. Big Data is an emerging paradigm that focuses on architecting analytic applications that can harness the petabytes of complex information flowing in from social media and other new and traditional sources [Jam11]. This is considered by Forbes as the biggest game-changing opportunity for marketing and sales since the Internet went mainstream 20 years ago [GPS13]. The biggest players have recognized the possibilities. Financial institutions alone, are expected to spend $3.2 billion in 2017 on Customer Analytics [Kee15]. From tech firms like Google, LinkedIn, Amazon, and Uber, to legacy firms like the Gap, Zurich Insurance, General Motors, Clorox, and AIG, organizations in every industry are turning to data science for business critical insights [Bur15].

Businesses today must excel at transforming the data derived from customer interactions into insights that fuel improved customer understanding [BSB14], but most smaller companies lack the know-how, can't afford a ground-up implementation, or try and fail. Report by CBS estimates that Danish companies loose at least 1 billion kr. a year, due to failed CRM projects. In 2013 only 15% of Danish companies had an actual customer strategy, and 25% of them admitted to having no strategy at all [Jac13]. Gartner and Forrester identify the three biggest obstacles to building a successful CRM strategy, founded on Big Data customer analytics: fragmentation of available data, lack of sophistication of available tools [MSM14], and most importantly, shortage of talent [PSM+15]. McKinsey projects a shortage of 140,000 to 190,000 "deep analytic talent positions" plus a deficiency of 1.5 million analysts and analytics managers by 2018 in the US [MCB+11]. In Denmark alone, there will be an estimated shortage of 13.500 engineers by 2025, with big data skills among top 5 required capabilities, as estimated by Dansk Industri [Cor15]. Microsoft in Denmark together with DTU Compute are uniquely positioned to address these obstacles, and democratize customer analytics, for hundreds of Danish companies and over 40.000 worldwide. Through software innovation, we aim to reduce their reliance on elusive talent, cut the costs of integrating data sources, and train local specialists through internships.

The CBS study emphasizes the importance of understanding different segments of existing and prospective customers [Jac13]. Established methods such as behavioral customer segmentation and churn analysis can inform acquisition, retention, and loyalty strategies, but more accurate predictions of complex customer behaviors require multidimensional analysis. In our approach we blend proven methods like machine learning, social media analysis and dynamic network analysis (which DTU Compute specializes in) and create a hybrid approach not yet explored, by either competition or academia. Social media analytics yields brand sentiment, while social network analytics provides peer influence; both methods can increase the accuracy of a customer lifetime model through additional dimensions of analysis [MSM14]. In statistics and machine learning this method is known as ensemble modeling, where multiple models are combined to gain better predictive performance than could be achieved with the individual source models. Between 2006 and 2009, various contenders and the eventual winners of the $1 million Netflix prize used ensemble learning to improve the performance of the movie recommendation algorithm by more than 10% [Sri14].

Microsoft is committed to enable any business to transform itself through the power of data. A new suite, called Cortana Analytics, is designed to "democratize big data" says Microsoft CEO Satya Nadella [Nad15]. Microsoft's donation of $1 billion worth of Cloud computing services, including analytics, was announced in January 2016 [Win16], further emphasizing our commitment to

collaborations with Academia (such as DTU Compute). In the Cloud + Enterprise Division, which includes our Danish development center, we consider intelligent features leveraging big data as the biggest opportunity to differentiate our products. In Denmark, we have, over the last 3 years, dedicated a team of 21 engineers to developing Social Analytics features for the Microsoft Social Engagement (MSE) product. MSE along with Dynamics CRM, developed in Lyngby, is our delivery vehicle for these features, to 355[1] of the largest companies in Denmark, and over 40.000 worldwide.

Given the scale of potential impact, and alignment with Microsoft's overall vision, for us, it is no longer a question "if" or "when" to democratize customer analytics - but "where". Our candidate for your program received his Master's degree at the age of 22, is published and certified in relevant fields, has 7 years of engineering experience at Microsoft and CERN, and 2 years in Social Analytics. We have a close relationship with DTU, welcome 15 students for practice periods, 20 hours a week, and are willing to train more, within this project. The areas of expertise within the Cognitive Systems Group at DTU Compute create an ideal match for this venture. In the end though, it's the existence and support of a public Innovation's Fund, which makes a deciding argument for this project to be conducted in Denmark.

[1] number of Danish enterprise grade accounts implementing Dynamics CRM, including 24 municipalities and 70% of labor unions

Bibliography

[ALK16] Khaled Almgren, Jeongkyu Lee, and Minkyu Kim. Prediction of image popularity over time on social media networks. *2016 Annual Connecticut Conference on Industrial Electronics, Technology and Automation, CT-IETA 2016*, pages 1–6, 2016.

[AOS+16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[AP17] Robert D. Austin and Gary P. Pisano. Neurodiversity Is a Competitive Advantage. *Harvard Business Review*, 2017.

[ASH13] Mohamed Ahmed, Stella Spagna, and Felipe Huici. A Peek into the Future : Predicting the Evolution of Popularity in User Generated Content. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013.

[AW11a] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, sep 2011.

[AW11b] Sinan Aral and Dylan Walker. Identifying social influence in networks using randomized experiments. *IEEE Intelligent Systems*, 26(5):91–96, sep 2011.

[Axe97] Robert Axelrod. The Dissemination of Culture. *Journal of Conflict Resolution*, 41(2):203–226, apr 1997.

[BAK17] Nilayan Bhattacharya, I. Budak Arpinar, and Ugur Kursuncu. Real Time Evaluation of Quality of Search Terms during Query

Expansion for Streaming Text Data Using Velocity and Relevance. In *Proceedings - IEEE 11th International Conference on Semantic Computing, ICSC 2017*, pages 280–281. Institute of Electrical and Electronics Engineers Inc., mar 2017.

[Bar93] John Perry Barlow. Selling Wine Without Bottles: The Economy of Mind on the Global Net | Electronic Frontier Foundation. *EFF*, 1993.

[BB15] Roberts Buchanan and Anita Buchanan. CompuCom® Survey: Customer Experience Top Reason Companies Use Big Data Analytics | Business Wire, 2015.

[BBC+19] Theo Bertram, Elie Bursztein, Stephanie Caro, Hubert Chao, Rutledge Chin Feman, Peter Fleischer, Albin Gustafsson, Jess Hemerly, Chris Hibbert, Luca Invernizzi, et al. Five years of the right to be forgotten. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 959–972, 2019.

[BBM+16] Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for deep neural network architectures. In *Information Science and Applications (ICISA) 2016*, pages 913–922. Springer, 2016.

[Bey19] Brian Beyersdorf. Regulating the Most Accessible Marketplace of Ideas in History: Disclosure Requirements in Online Political Advertisements after the 2016 Election. *California Law Review*, 107, 2019.

[BFSO84] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[BGU16] Sneha Binani, Ajinkya Gutti, and Shivam Upadhyay. Sql vs. nosql vs. newsql-a comparative study. *database*, 6(1):1–4, 2016.

[BJ19] Keith James Bybee and Laura Jenkins. Free Speech, Free Press, and Fake News: What If the Marketplace of Ideas Isn't About Identifying Truth? *SSRN Electronic Journal*, jan 2019.

[BKA09] Eytan Bakshy, Brian Karrer, and Lada A. Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 325–334, 2009.

[BL12] Mark A Beyer and Douglas Laney. The importance of 'big data': a definition. *Stamford, CT: Gartner*, pages 2014–2018, 2012.

[BLSA17] Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. Unstructured point cloud semantic labeling using deep segmentation networks. *3DOR*, 2:7, 2017.

[BNQ19] Seyed Mojtaba Hosseini Bamakan, Ildar Nurgaliev, and Qiang Qu. Opinion leader detection: A methodological review. *Expert Systems with Applications*, 115:200–222, 2019.

[BOJC16] Gema Bello-Orgaz, Jason J. Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, mar 2016.

[BP08] Abraham Bagherjeiran and Rajesh Parekh. Combining behavioral and social network data for online advertising. In *2008 IEEE International Conference on Data Mining Workshops*, pages 837–846. IEEE, 2008.

[Bre15] Eric A Brewer. Kubernetes and the path to cloud native. In *Proceedings of the Sixth ACM Symposium on Cloud Computing*, pages 167–167, 2015.

[BSB14] Michael Barnes, Srividya Sridharan, and Zhi-Ying Barry. Asia Pacific Companies Embrace Customer Analytics, Part 1. Technical report, Forrester, 2014.

[BSG14] Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. *Conference on Human Factors in Computing Systems - Proceedings*, pages 965–974, 2014.

[BT16] Hendra Bunyamin and Tomas Tunys. A comparison of retweet prediction approaches: the superiority of random forest learning method. *Telkonika (Telecommun Comput Electron Control)*, 14(3):1052–1058, 2016.

[Bto...17] V Badrinarayanan, A Kendall IEEE transactions on ..., and Undefined 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *ieeexplore.ieee.org*, 2017.

[Bue16] Claudio Celis Bueno. The attention economy: Labour, time and power in cognitive capitalism. 2016.

[Bur15] Linda Burtch. Burtch Works 2015 Predictions - Analytics/Data Science Hiring, 2015.

[BY14] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 1:316–334, 2014.

[CAD+14] Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon Kleinberg, and Jure Leskovec. Can Cascades be Predicted? mar 2014.

[CFFV16] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016.

[CFM15] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. The production of information in the attention economy. *Scientific Reports*, 5(1):1–6, may 2015.

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM 10*, 2010.

[CHJ+16] Samuel Carton, Jennifer Helsby, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Joe Walsh, Crystal Cody, CPT Estella Patterson, Lauren Haynes, and Rayid Ghani. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 67–76, 2016.

[Cho10] C Chodorow. Introduction to mongodb. In *Free and Open Source Software Developers European Meeting (FOSDEM)*, 2010.

[CJS18] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.

[CKK14] Ray M. Chang, Robert J. Kauffman, and Youngok Kwon. Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63:67–80, jul 2014.

[CKXM19] Guandan Chen, Qingchao Kong, Nan Xu, and Wenji Mao. NPP: A neural popularity prediction model for social media content. *Neurocomputing*, 333:221–230, 2019.

[CLZ+19] Junhong Chen, Dayong Liang, Zhanmo Zhu, Xiaojing Zhou, Zihan Ye, and Xiuyun Mo. Social media popularity prediction based on visual-textual features with XGboost. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, pages 2692–2696, 2019.

[CMS12] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[CMS15] Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. Latent factors of visual popularity prediction. *ICMR 2015 - Proceedings of the 2015 ACM International Conference on Multimedia Retrieval*, pages 195–202, 2015.

[Coh88] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

[COM13] Ethem F. Can, Hüseyin Oktay, and R. Manmatha. Predicting retweet count using visual cues. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, 2013.

[Cor15] Cordis. Support to the Grand Coalition for ICT jobs: Conclusion and recommendations from local and regional networking activities. Technical report, European Commission, jul 2015.

[CPK+14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[CS08] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[Dat19] DataSift. Fairhair.ai Data Platform reference. `https://developers.fairhair.ai/docs/data-platform-overview`, 2019. Accessed: 2020-03-20.

[Daw76] Richard Dawkins. *The selfish gene*. 1976.

[DB01] Thomas H Davenport and John C Beck. *The attention economy: Understanding the new currency of business*. Harvard Business Press, 2001.

[DCL+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical report, 2018.

[DG10] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: a flexible data processing tool. *Communications of the ACM*, 53(1):72–77, 2010.

[DMCF15] Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 2015.

[DMW19] Keyan Ding, Kede Ma, and Shiqi Wang. Intrinsic image popularity assessment. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, (October):1979–1987, 2019.

[Dom05] Pedro Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1):80–82, 2005.

[DP15] Arturo Deza and Devi Parikh. Understanding image virality. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1818–1826, 2015.

[DSZ16] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

[DVGK14] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

[DVK17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[DWW19] Keyan Ding, Ronggang Wang, and Shiqi Wang. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, pages 2682–2686, 2019.

[Dze00] Martin Dzelzainis. John milton, areopagitica. *A Companion to Literature from Milton to Blake*, pages 151–8, 2000.

[EMR+19] Soheil Eshgi, Setareh Maghsudi, Valerio Restocchi, Sebastian Stein, and Leandros Tassiulas. Efficient influence maximization under network uncertainty. In *INFOCOM 2019 Workshop proceedings*, 2 2019.

[ESTA14] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

[Fal07] Josef Falkinger. Attention economies. *Journal of Economic Theory*, 133(1):266–294, mar 2007.

[FB13] Wei Fan and Albert Bifet. Mining big data. *ACM SIGKDD Explorations Newsletter*, 14(2):1, apr 2013.

[FDS16] Syeda Nadia Firdaus, Chen Ding, and Alireza Sadeghian. Retweet prediction considering user's difference as an author and retweeter. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 2016.

[Fel13] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 2013.

[Fow18] Martin Fowler. *Refactoring: improving the design of existing code.* Addison-Wesley Professional, 2018.

[Fra99] Georg Franck. The Economy of Attention. *Telepolis*, 1999.

[Fra19] Georg Franck. The economy of attention. *Journal of Sociology*, 55(1):8–19, mar 2019.

[FVD+16] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, jun 2016.

[FZ15] Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5, 2015.

[GA15] Satish Gopalani and Rohan Arora. Comparing apache spark and map reduce with performance analysis using k-means. *International Journal of Computer Applications*, 113:8–11, 03 2015.

[GF17] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[GH15] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 2015.

[GHTA18] Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. Social media big data analytics: A survey. *Computers in Human Behavior*, 2018.

[GIL12] Anindya Ghose, Panagiotis G. Ipeirotis, and Beibei Li. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, 31(3):493–520, may 2012.

[GJ18] Mingxin Gan and Rui Jiang. FLOWER: Fusing global and local associations towards personalized social recommendation. *Future Generation Computer Systems*, 2018.

[GO19] Mehmetcan Gayberi and Sule Gunduz Oguducu. Popularity prediction of posts in social networks based on user, post and image features. *11th International Conference on Management of Digital EcoSystems, MEDES 2019*, pages 9–15, 2019.

[Gol97] Michael H. Goldhaber. The attention economy and the net. *First Monday*, 2(4), apr 1997.

[Gol06] Michael Goldhaber. The value of openness in an attention economy. *First Monday*, 11(6), jun 2006.

[GPS13] Jonathan Gordon, Jesko Perry, and Dennis Spillecke. Big Data, Analytics And The Future Of Marketing And Sales, 2013.

[GPV11] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE*, 6(8):e22656, aug 2011.

[Gua18] José Rolando Guay Paz. Introduction to Azure Cosmos DB. In *Microsoft Azure Cosmos DB Revealed*, pages 1–23. Apress, Berkeley, CA, 2018.

[GUB+15] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih Fu Chang. Image popularity prediction in social media using sentiment and context features. *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*, pages 907–910, 2015.

[GWDC16] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. Quote rts on twitter: Usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 200–204, New York, NY, USA, 2016. ACM.

[GWZW16] Dong Guo, Wei Wang, Guosun Zeng, and Zerong Wei. Microservices architecture based cloudware deployment platform for service computing. In *2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)*, pages 358–363. IEEE, 2016.

[HAN+11] Lars Kai Hansen, Adam Arvidsson, Finn Aarup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news - Affect and virality in twitter. In *Communications in Computer and Information Science*, 2011.

[HBA15] Ilkyu Ha, Bonghyun Back, and Byoungchul Ahn. Mapreduce functions to analyze sentiment information from social big data. *International Journal of Distributed Sensor Networks*, 11(6):417502, 2015.

[HBPK17] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? dec 2017.

[HCYH17] Shintami Chusnul Hidayati, Yi Ling Chen, Chao Lung Yang, and Kai Lung Hua. Popularity meter: An influence- and aesthetics-aware social media popularity predictor. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, pages 1918–1923, 2017.

[HDWF+17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[Hea10] Alison Hearn. Structuring feeling: Web 2.0, online ranking and rating, and the digital 'reputation' economy. Technical report, 2010.

[HHWY19] Ziliang He, Zijian He, Jiahong Wu, and Zhenguo Yang. Feature construction for posts and users combined with lightgBM for social media popularity prediction. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, pages 2672–2676, 2019.

[HKS71] Speaker A Herbert Simon, Discussants W Karl Deutsch, and Martin Shubik. Designing Organizations For An In Formation-Rich World. Technical report, 1971.

[HLK+19] Chih Chung Hsu, Jun Yi Lee, Li Wei Kang, Zhong Xuan Zhang, Chia Yen Lee, and Shao Min Wu. Popularity prediction of social media based on multi-modal feature mining. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, pages 2687–2691, 2019.

[Hor65] John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[HS15a] Kevan Harris and Ben Scully. A hidden counter-movement? precarity, politics, and social protection before and beyond the neoliberal era. *Theory and Society*, 44(5):415–444, 2015.

[HS15b] Daniel E Ho and Frederick Schauer. Testing the marketplace of ideas. *NYUL Rev.*, 90:1160, 2015.

[HTE10] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in Twitter. In *HT'10 - Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 173–177, New York, New York, USA, 2010. ACM Press.

[HYA+15] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, 47:98–115, 2015.

[HYA16] Jiani Hu, Toshihiko Yamasaki, and Kiyoharu Aizawa. Multimodal learning for image popularity prediction on social media. *2016 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2016*, pages 1–2, 2016.

[HZRS16a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 770–778. IEEE Computer Society, dec 2016.

[HZRS16b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[IKT12] Katsuhiko Ishiguro, Akisato Kimura, and Koh Takeuchi. Towards automatic image understanding and mining via social curation. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2012.

[Jac13] Per Østergaard Jacobsen. Virksomhedernes Kunderelationer 2013 CRM i danske virksomheder. Technical report, Copenhagen Business School, Copenhagen, 2013.

[Jam11] James G. Kobielus. How Social CRM Benefits From Big Data. Technical report, Forrester, 2011.

[JH11] Kasper Winther Jorgensen and Lars Kai Hansen. Model selection for gaussian kernel pca denoising. *IEEE transactions on neural networks and learning systems*, 23(1):163–168, 2011.

[Jos20] Ameet V Joshi. Azure machine learning. In *Machine Learning and Artificial Intelligence*, pages 207–220. Springer, 2020.

[KAEM13] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, and William Money. Big data: Issues and challenges moving forward. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2013.

[KBHG14] Oleksii Kononenko, Olga Baysal, Reid Holmes, and Michael W Godfrey. Mining modern repositories with elasticsearch. In *Proceedings of the 11th working conference on mining software repositories*, pages 328–331, 2014.

[KDH14] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? pages 867–876, 2014.

[Kee15] Charles Keenan. Big Data and Predictive Analytics: A Big Deal, Indeed. *ABA Banking Journal*, 2015.

[Kel99] Kevin Kelly. *New rules for the new economy: 10 radical strategies for a connected world*. Penguin, 1999.

[KH20] Damian Konrad Kowalczyk and Lars Kai Hansen. The complexity of social media response: Statistical evidence for one-dimensional engagement signal in twitter. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, 2020.

[KJKW18] Hwalbin Kim, S. Mo Jang, Sei-Hill Kim, and Anan Wan. Evaluating Sampling Methods for Content Analysis of Twitter Data. *Social Media + Society*, 4(2):205630511877283, apr 2018.

[KL19] Damian Konrad Kowalczyk and Jan Larsen. Scalable Privacy-Compliant Virality Prediction on Twitter. In *Proceedings of AffCon 2019 @ AAAI*, volume 2328, pages 12–27, 2019. (Best Paper Award).

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010.

[KLT+19] Peipei Kang, Zehang Lin, Shaohua Teng, Guipeng Zhang, Lingni Guo, and Wei Zhang. Catboost-based framework with additional user information for social media popularity prediction. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, pages 2677–2681, 2019.

[KMF+17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):3147–3155, 2017.

[KMT19] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Metric space of collider events. *Phys. Rev. Lett.*, 123:041801, Jul 2019.

[KMW+17] Guolin Ke, Qi Meng, Taifeng Wang, Wei Chen, Weidong Ma, Tie-Yan Liu, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017.

[KSN20] Akshi Kumar, Saurabh Raj Sangwan, and Anand Nayyar. Multimedia social big data: Mining. In *Multimedia Big Data Computing for IoT Applications*, pages 289–321. Springer, 2020.

[Lan01] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

[LBHW20] Qin Liu, Md Zakirul Alam Bhuiyan, Jiankun Hu, and Jie Wu. Preface: Security & privacy in social big data, 2020.

[LD19] Weidong Liao and Jesse Draper. Cloud computing and docker containerization: A survey. *Proceedings of the West Virginia Academy of Science*, 91(1), 2019.

[Lea16] Leansoft. A big, fast and persistent queue based on memory mapped file. `https://github.com/bulldog2011/bigqueue`, 2016. Accessed: 2020-03-20.

[LEL18] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles, 2018.

[LHN18] Dokyun Lee, Kartik Hosanagar, and Harikesh S Nair. Advertising content and consumer engagement on social media: evidence from facebook. *Management Science*, 64(11):5105–5131, 2018.

[Lip06] Stan Lipovetsky. Entropy criterion in logistic regression and shapley value of predictors. *Journal of Modern Applied Statistical Methods*, 5(1):9, 2006.

[LL17a] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-December(Section 2):4766–4775, 2017.

[LL17b] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[LMB+14] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014.

[Lom06] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

[LS20] In Lee and Yong Jae Shin. Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, 63(2):157–170, 2020.

[LSMHL19] Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. Accelerating dynamics of collective attention. *Nature Communications*, 10(1):1759, dec 2019.

[LWJ+20] Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, et al. The microsoft toolkit of multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:2002.07972*, 2020.

[MB17] Garrett McGrath and Paul R Brenner. Serverless computing: Design, implementation, and performance. In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 405–410. IEEE, 2017.

[MBY+15] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. MLlib: Machine Learning in Apache Spark. may 2015.

[MBY+16] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. Mllib: Machine learning in apache spark. *J. Mach. Learn. Res.*, 17(1):1235–1241, January 2016.

[MCA+13] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H Eugene Stanley, and Tobias Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific reports*, 3:1801, 2013.

[MCB+11] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity | McKinsey. Technical report, McKinsey Global Institute, 2011.

[MFGZ18] Maciej Malawski, Kamil Figiela, Adam Gajek, and Adam Zima. Benchmarking heterogeneous cloud functions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10659 LNCS, pages 415–426. Springer Verlag, 2018.

[MG18] Suren Machiraju and Suraj Gaurav. *DevOps for Azure Applications*. Springer, 2018.

[MGTW12] Uskali Mäki, Dov M. Gabbay, Paul Thagard, and John Woods. *Philosophy of Economics*. Elsevier, 2012.

[Mic17] Microsoft. Cognitive Services Text Analytics API reference. `https://westus.dev.cognitive.microsoft.com/docs/services/TextAnalytics-v3-0-Preview-1/operations/Sentiment`, 2017. Accessed: 2020-03-20.

[Mic19a] Microsoft. Bing Maps Locations API reference. `https://docs.microsoft.com/en-us/bingmaps/rest-services/locations/`, 2019. Accessed: 2020-03-20.

[Mic19b] Microsoft. Cognitive Services Translator API reference. `https://docs.microsoft.com/en-us/azure/cognitive-services/translator/reference/v3-0-detect`, 2019. Accessed: 2020-03-20.

[Mic19c] Microsoft. Microsoft Social Engagement release notes. `https://docs.microsoft.com/en-us/dynamics365/customer-engagement/social-engagement/what-s-new-in-microsoft-social-engagement`, 2019. Accessed: 2019-04-20.

[Mil19] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[MKS+13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[MLCM13] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3(1):1–7, 2013.

[MMJ14] Philip J. McParlane, Yashar Moshfeghi, and Joemon M. Jose. "Nobody comes here anymore, it's too crowded"; predicting image

popularity on Flickr. *ICMR 2014 - Proceedings of the ACM International Conference on Multimedia Retrieval 2014*, pages 385–391, 2014.

[Mol19] Christoph Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[MPW18] Masoud Mazloom, Iliana Pappi, and Marcel Worring. Category Specific Post Popularity Prediction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10704 LNCS:594–607, 2018.

[MRR+16] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. Multimodal popularity prediction of brand-related social media posts. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, pages 197–201, 2016.

[MSC13] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[MSM14] Jason McNellis, Sridharan Srividya, and Rebecca McAdams. The State Of Customer Analytics: Majority Of Firms Lack Sophistication. Technical report, Forrester, 2014.

[Nad15] Satya Nadella. Worldwide Partner Conference 2015 - Stories, jul 2015.

[NPPZ18] Paolo Nesi, Gianni Pantaleo, Irene Paoli, and Imad Zaza. Assessing the reTweet proneness of tweets: predictive models for retweeting. *Multimedia Tools and Applications*, 2018.

[NR10] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, jun 2010.

[OFB19] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. Predicting Social Image Popularity Dynamics at Time Zero. *IEEE Access*, 7:171691–171706, 2019.

[OMW+17] Gijs Overgoor, Masoud Mazloom, Marcel Worring, Robert Rietveld, and Willemijn Van Dolen. A spatio-temporal category representation for brand popularity prediction. *ICMR 2017 - Proceedings of the 2017 ACM International Conference on Multimedia Retrieval*, pages 233–241, 2017.

[Ott13] Clemens Otte. Safe and interpretable machine learning: A methodological review. In *Computational intelligence in intelligent data analysis*, pages 111–122. Springer, 2013.

[PAP+13] Fabio Pezzoni, Jisun An, Andrea Passarella, Jon Crowcroft, and Marco Conti. Why do I retweet it? An information propagation model for microblogs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013.

[Par11] Eli Pariser. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Technical report, 2011.

[PBC12] Jomon Aliyas Paul, Hope M Baker, and Justin Daniel Cochran. Effect of online social networking on student academic performance. *Computers in Human Behavior*, 28(6):2117–2127, 2012.

[PDB13] Robert Palovics, Balint Daroczy, and Andras A. Benczur. Temporal prediction of retweet count. In *4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013 - Proceedings*, 2013.

[PGK+07] A Petrucci, M Gulmini, JC Kim, I Suzuki, M Klute, S Erhan, V O'dell, S Murray, C Schwick, J Varela, et al. The run control and monitoring system of the cms experiment. Technical report, 2007.

[Pim90] Martti Pimiä. Compact muon solenoid. Technical report, CERN, 1990.

[PSM+15] Brandon Purcell, Srividya Sridharan, Jason McNellis, Shar VanBoskirk, and Olivia French. A Stopgap For Data Science Scarcity. Technical report, Forrester, 2015.

[PYM18] Sancheng Peng, Shui Yu, and Peter Mueller. Social networking big data: Opportunities, solutions, and challenges. *Future Generation Computer Systems*, 86:1456–1458, sep 2018.

[PZP+11] Huan Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. Retweet modeling using conditional random fielDs. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2011.

[QFS+17] Xiaoyan Qiu, Diego F. M. Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. Limited individual attention and online virality of low-quality information. *Nature*

Human Behaviour, 1(7):0132, July 2017. (Retraction published 07 January 2019, Nature Human Behaviour, 3(1), 102-102).

[RCC20] Juan Francisco Robles, Manuel Chica, and Oscar Cordon. Evolutionary multiobjective optimization to target social network influentials in viral marketing. *Expert Systems with Applications*, page 113183, 2020.

[RCM+11] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. Detecting and tracking political abuse in social media. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[RDGF16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. Technical report, 2016.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[RF18] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. 2018.

[RKH20] Christoffer Riis, Damian Konrad Kowalczyk, and Lars Kai Hansen. On the limits to multi-modal popularity prediction on instagram – a new robust, efficient and explainable baseline. 2020. (Submitted).

[RP18] José Rolando and Guay Paz. Microsoft Azure Cosmos DB Revealed A Multi-Model Database Designed for the Cloud-Building globally distributed mission-critical applications. 2018.

[RR18] Rob Reagan and Rob Reagan. Cosmos DB. In *Web Applications on Azure*, pages 187–255. Apress, 2018.

[RSD+17] Raghu Ramakrishnan, Baskar Sridharan, John R Douceur, Pavan Kasturi, Balaji Krishnamachari-Sampath, Karthick Krishnamoorthy, Peng Li, Mitica Manu, Spiro Michaylov, Rogério Ramos, et al. Azure data lake store: a hyperscale distributed file service for big data analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 51–63, 2017.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[RvDMW20] Robert Rietveld, Willemijn van Dolen, Masoud Mazloom, and Marcel Worring. What You Feel, Is What You Like Influence of Message Appeals on Customer Engagement on Instagram. *Journal of Interactive Marketing*, 49:20–53, 2020.

[RW17] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[Sam19] Wojciech Samek. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[SCLW17] Mohammad Javad Shafiee, Brendan Chywl, Francis Li, and Alexander Wong. Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video. *Journal of Computational Vision and Imaging Systems*, 3(1), sep 2017.

[Sco11] John Scott. Social physics and social networks. *The SAGE handbook of social network analysis*, pages 55–66, 2011.

[SGK17] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[She16] L. Jaba Sheela. A Review of Sentiment Analysis in Twitter Data Using Hadoop. *International Journal of Database Theory and Application*, 2016.

[SHG+15] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511, 2015.

[SHM+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[SNP12] Alfred Spector, Peter Norvig, and Slav Petrov. Google's hybrid approach to research. *Communications of the ACM*, 55(7):34–37, 2012.

[Sor15] Thompson Ed Sorofman Jake. Customer Experience Is the New Competitive Battlefield. Technical report, Gartner, 2015.

[SP18] Androniki Sapountzi and Kostas E. Psannis. Social networking data analysis tools & challenges. *Future Generation Computer Systems*, 2018.

[Sri14] Srividya Sridharan. TechRadar^{TM}: Customer Analytics Methods, Q1 2014. Technical report, 2014.

[SRN+18a] Alexander Spangher, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. Analysis of strategy and spread of russia-sponsored content in the us in 2017. *arXiv preprint arXiv:1810.10033*, 2018.

[SRN+18b] Alexander Spangher, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. Analysis of strategy and spread of russia-sponsored content in the us in 2017. *ArXiv*, abs/1810.10033, 2018.

[STE13] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. In *Advances in neural information processing systems*, pages 2553–2561, 2013.

[Sun14] CR Sunstein. On rumors: How falsehoods spread, why we believe them, and what can be done. 2014.

[SVZ13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[SWM17] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. aug 2017.

[TAdAF14] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida. From popularity prediction to ranking online news. *Social Network Analysis and Mining*, 4(1):1–12, jan 2014.

[TADF12] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. Ranking news articles based on popularity prediction. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pages 106–110, 2012.

[TBDH19] Robert Thomson, Halil Bisgin, Christopher Dancy, and Ayaz Hyder. *Social, Cultural, and Behavioral Modeling*. Springer, 2019.

[Ter12] Tiziana Terranova. Attention, economy and the brain. *Culture Machine*, 13, 2012.

[TL19a] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019.

[TL19b] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:10691–10700, may 2019.

[TLP14] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[TSS17] Rebekah Tromble, Andreas Storz, and Daniela Stockmann. We don't know what we don't know: When and how the use of twitter's public apis biases scientific inference. *Available at SSRN 3079927*, 2017.

[Twi19] Twitter. Developer API Docs: Tweet objects. `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object`, 2019. Accessed: 2020-03-20.

[VA19] María Vega García and José L Aznarte. Shapley additive explanations for NO 2 forecasting. 2019.

[VBZ+16] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3):554–559, jan 2016.

[vK19] Robert van Krieken. Georg franck's 'the economy of attention': Mental capitalism and the struggle for attention. *Journal of Sociology*, 55(1):3–7, 2019.

[Wal58] Walter D. Fisher. On Grouping For Maximum Homogeneity. *American Statistical Association Journal*, 1958.

[WBF18] Ke Wang, Mohit Bansal, and Jan Michael Frahm. Retweet wars: Tweet popularity prediction via dynamic multimodal regression. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, 2018-Janua:1842–1851, 2018.

[WCZM16] Bo Wu, Wen Huang Cheng, Yongdong Zhang, and Tao Mei. Time matters: Multi-scale temporalization of social media popularity.

MM 2016 - Proceedings of the 2016 ACM Multimedia Conference, pages 1336–1344, 2016.

[WFVM12] L Weng, A Flammini, A Vespignani, and F Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2:335, 2012.

[Win16] Nick Wingfield. Microsoft to Donate $1 Billion in Cloud Services to Nonprofits and Researchers - The New York Times, jan 2016.

[WJC+15] Helena Webb, Marina Jirotka, Bernd Carsten Stahl, William Housley, Adam Edwards, Matthew Williams, Rob Procter, Omer Rana, and Pete Burnap. 'Digital wildfires': A challenge to the governance of social media? In *Proceedings of the 2015 ACM Web Science Conference*. Association for Computing Machinery, Inc, jun 2015.

[WS08] Jonathan Wareham and Thorkil Sonne. Harnessing the Power of Autism Spectrum Disorder (Innovations Case Narrative: Specialisterne) . *Innovations: Technology, Governance, Globalization*, 3(1):11–27, jan 2008.

[WS15] Bo Wu and Haiying Shen. Analyzing and predicting news popularity on Twitter. *International Journal of Information Management*, 2015.

[Wu17] Tim Wu. The attention merchants: The epic scramble to get inside our heads. 2017.

[WZS+16] Yakun Wang, Zhongbao Zhang, Sen Su, Cheng Chang, and Muhammad Azam Zia. Topic-Level Influencers Identification in the Microblog Sphere. 0:4–5, 2016.

[XDG+14] Reynold Xin, Parviz Deyhim, Ali Ghodsi, Xiangrui Meng, and Matei Zaharia. Graysort on apache spark by databricks. *GraySort Competition*, 2014.

[ZCZ+18] Zhongping Zhang, Tianlang Chen, Zheng Zhou, Jiaxin Li, and Jiebo Luo. How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pages 2383–2392, 2018.

[ZEH+15] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. *CoRR*, abs/1506.02594, 2015.

[Zhe14] A Zheng. The challenges of building machine learning tools for the masses. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[ZHvGS10] Tauhid R Zaman, Ralf Herbrich, Jurgen van Gael, and David Stern. Predicting Information Spreading in Twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS 2010*, 2010.

[ZJ19] Yihong Zhang and Adam Jatowt. Image tweet popularity prediction with convolutional neural network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11437 LNCS, pages 803–809. Springer Verlag, 2019.

[ZK16] Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *British Machine Vision Conference 2016, BMVC 2016*, volume 2016-September, pages 87.1–87.12. British Machine Vision Conference, BMVC, may 2016.

[ZLK+18a] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

[ZLK+18b] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, jun 2018.

[ZLX+14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning Deep Features for Scene Recognition using Places Database - Supplementary Materials. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, 1:487–495, 2014.

[ZSH17] Huan Zhang, Si Si, and Cho-Jui Hsieh. GPU-acceleration for Large-scale Tree Boosting. jun 2017.

[ZSY18] Alireza Zohourian, Hedieh Sajedi, and Arefeh Yavary. Popularity prediction of images and videos on Instagram. *2018 4th International Conference on Web Research, ICWR 2018*, pages 111–117, 2018.

[ZXW+16] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. Apache spark:

a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.