

What sets Verified Users apart? Insights, Analysis and Prediction of Verified Users on Twitter

by

Indraneil Paul paul, Abhinav Khattar, Shaan Chopra, Ponnurangam kumaraguru, Manish Gupta

in

*11th ACM Conference on Web Science
(WebSci-2019)*

Boston, MA, USA

Report No: IIIT/TR/2019/-1



Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2019

What sets Verified Users apart? Insights, Analysis and Prediction of Verified Users on Twitter

Indraneil Paul
IIIT Hyderabad
indraneil.paul@research.iiit.ac.in

Abhinav Khattar
IIIT Delhi
abhinav15120@iiitd.ac.in

Shaan Chopra
IIIT Delhi
shaan15090@iiitd.ac.in

Ponnurangam Kumaraguru*
IIIT Delhi
pk@iiitd.ac.in

Manish Gupta**
IIIT Hyderabad
manish.gupta@iiit.ac.in

ABSTRACT

Social network and publishing platforms, such as Twitter, support the concept of a secret proprietary *verification* process, for handles they deem worthy of platform-wide public interest. In line with significant prior work which suggests that possessing such a status symbolizes enhanced credibility in the eyes of the platform audience, a verified badge is clearly coveted among public figures and brands. What are less obvious are the inner workings of the verification process and what being verified represents. This lack of clarity, coupled with the flak that Twitter received by extending aforementioned status to political extremists in 2017, backed Twitter into publicly admitting that the process and what the status represented needed to be rethought.

With this in mind, we seek to unravel the aspects of a user's profile which likely engender or preclude verification. The aim of the paper is two-fold: First, we test if discerning the verification status of a handle from profile metadata and content features is feasible. Second, we unravel the features which have the greatest bearing on a handle's verification status. We collected a dataset consisting of profile metadata of all 231,235 verified English-speaking users (as of July 2018), a control sample of 175,930 non-verified English-speaking users and all their 494 million tweets over a one year collection period. Our proposed models are able to reliably identify verification status (Area under curve AUC > 99%). We show that number of public list memberships, presence of neutral sentiment in tweets and an authoritative language style are the most pertinent predictors of verification status.

To the best of our knowledge, this work represents the first attempt at discerning and classifying verification worthy users on Twitter.

* This work was partially done by the author while in sabbatical at IIIT Hyderabad (pk.guru@iiit.ac.in).

** The author is also a Principal Applied Researcher at Microsoft India (gmanish@microsoft.com).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6202-3/19/06...\$15.00
<https://doi.org/10.1145/3292522.3326026>

CCS CONCEPTS

• **Information systems** → **Social networks**; *Relevance assessment*; Content analysis and feature selection; • **Networks** → **Social media networks**; **Online social networks**.



KEYWORDS

Twitter, Social Influence, Verified Users

ACM Reference Format:

Indraneil Paul, Abhinav Khattar, Shaan Chopra, Ponnurangam Kumaraguru, and Manish Gupta. 2019. What sets Verified Users apart? Insights, Analysis and Prediction of Verified Users on Twitter. In *11th ACM Conference on Web Science (WebSci '19)*, June 30-July 3, 2019, Boston, MA, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3292522.3326026>

1 INTRODUCTION

The increased relevance of social media in our daily life has been accompanied by an exigent demand for a means to affirm the authenticity and authority of content sources. This challenge becomes even more apparent during the dissemination of real-time or breaking news, whose arrival on such platforms often precedes eventual traditional media reportage [19, 38]. In line with this need, major social networks such as Twitter, Facebook and Instagram have incorporated a verification process to authenticate handles they deem important enough to be worth impersonating. Usually conferred to accounts of well-known public personalities and businesses, *verified accounts*¹ are indicated with a badge next to the screen name (e.g.,  on Twitter and  on Facebook). Twitter's verification policy [67] states that an account is verified if it belongs to a personality or business deemed to be of sufficient public interest in diverse fields, such as journalism, politics, sports, etc. However, the exact decision making process behind evaluating the strength of a user's case for verification remains a trade secret. This work attempts to unravel the likely factors that strengthen a user's case for verification by delving into the aspects of a user's Twitter presence, that most reliably predict platform verification.

1.1 Motivation

Our motivation behind this work was two-fold and is elaborated in the following text.

¹The exact term varies by platform, with other social networks using the term "Verified Profiles". However in the interest of consistency, all owner-authenticated accounts are referred to as *verified accounts*, and their owners as *verified users*.

Lack of procedural clarity and imputation of bias: Despite repeated statements by Twitter about verification not being equivalent to endorsement, aspects of the process – the rarity of the status and its prominent visual signalling [68] – have led users to conflate authenticity and credibility. This perception was confirmed in full public view when Twitter was backed into suspending its requests for verification in response to being accused of granting verified status to political extremists², with the insinuation being that the verified badge lent their otherwise extremist opinions a facade of mainstream credibility.

This however, engendered accusations of Twitter’s verification procedure harbouring a liberal bias. Multiple tweets imputing the same gave rise to the hashtag #VerifiedHate. Similar insinuations have been made by right-leaning Indian users of the platform in the lead up to the 2019 Indian General Elections under the hashtag #ProtestAgainstTwitter. These hitherto unfounded allegations of bias prompted us to delve deeper into understanding what may be driving the process and inferring whether these claims were justified or could the difference in status be explained away by less insidious factors relating to a user’s profile and content.

Positive perception and coveted nature: Despite having its detractors, the fact remains that a verified badge is highly coveted amongst public figures and influencers. This is with good reason as in spite of being intended as a mark of authenticity, prior work in social sciences and psychology points to verified badges conferring additional credibility to a handle’s posted tweets [11, 23, 51]. Psychological testing [24] has also revealed that the credibility of a message and its reception is influenced by its purported source and presentation rather than just its pertinence or credulity. Captology studies [21] indicate that widely endorsed information originating from a well-known source is easier to perceive as trustworthy and back up the former claim. This is pertinent as owners of verified accounts are usually well-known and their content is on average more frequently liked and retweeted than that of the generic Twittersphere [58, 63].

Adding to the desirability of exclusive visual indicators is the demanding nature of credibility assessment on Twitter. The imposed character limit and a minimal scope of visually customizing content, coupled with the feverish rate at which content is consumed – with users on average devoting a mere three seconds of attention per tweet [17] – makes users resort to heuristics to judge online content. There is substantial work on heuristic based models for online credibility evaluation [14, 29, 61]. Particularly relevant to this inquiry is the *endorsement heuristic*, which is associated with credibility conferred to it (e.g. a verified badge) and the *consistency heuristic*, which stems from endorsements by several authorities (e.g. a user verified in one platform is likely to be verified on others).

Unsurprisingly, a verified status is highly sought after by preeminent entities, as evidenced by the prevalence of get-verified-quick schemes such as promoted tweets from the now suspended account ‘@verified845’ [9, 65]. Our work attempts to obtain actionable insights into verification process, thus providing entities looking to get verified a means to strengthen their case.

1.2 Research Questions

The aforementioned motivating factors pose a few avenues of research inquiry that we attempt to answer, which are detailed below.

- RQ1:** Can the verification status of a user be predicted from profile metadata and tweet contents? If so what are the most reliably discriminative features?
- RQ2:** Do any inconsistencies exist between verified and non-verified users with respect to peripheral aspects like the choice and variety of topics they tweet about?

1.3 Contributions

Our contributions can be summarized as follows:

- We motivate and propose the problem of predicting verification status of a Twitter user.
- We detail a framework extracting a substantial set of features from data and meta-data about social media users, including friends, tweet content and sentiment, activity time series, and profile trajectories. We plan to make this dataset of 407,165 users and 494 million tweets, publicly available upon publication of the work.³
- Additionally, we factored in state-of-the-art bot detection analysis into our predictive model. We use these features to train highly-accurate models capable of discerning a user’s verified status. For a general user, we are able to provide a zero to one score representing their likelihood of being verified in Twitter.
- We report the most informative features in discriminating verified users from non-verified ones and also shed light on the manner in which the span and gamut of topic coverage between their tweets differs.

The rest of the paper is organized as follows. Section 2 details relevant prior work, hence putting our work in perspective. Section 3 elaborates our data acquisition methodology. In Sections 4 and 5, we conduct a comparative analysis between verified and non-verified users, addressing RQ1 and RQ2 respectively, and attempt to uncover features that can reliably classify them. We conclude with a brief summary in Section 6.

2 RELATED WORK

Previous studies have focused on measuring user impact in social networks. As user impact might be a critical factor in deciding who gets verified on Twitter [67], it is important to study how certain users in particular networks have more impact/influence as compared to the others. Cha et al. [13] studied the dynamics of influence on Twitter based on three key measures: in-degree, retweets, and user-mentions. They show that in-degree alone is not sufficient to measure the influence of a user on Twitter. Bakshy et al. [5] demonstrate that URLs from users who have been influential in the past tend to generate larger cascades on the Twitter follower graph. They also show that URLs considered more interesting and that kindle positive emotions, spread more. Canali et al. [10] identify key users on social networks who are important sources or targets for content disseminated online. They use a dimensionality-reduction based technique and conduct experiments with YouTube and Flickr

²<https://www.bbc.com/news/technology-41934831>

³<http://precog.iitd.edu.in/requester.php?dataset=twitterVerified19>

datasets to obtain results which outperform the existing solutions by 15%. The novelty of their approach is that they use attribute rich user profiles and not just stay limited to their network information. On the other hand, Lampos et al. [40] predict user impact on Twitter using features, such as user statistics and tweet content, that are under the control of the user. They experiment with both linear and non-linear prediction techniques and find that Gaussian Processes based models perform the best for the prediction task. Klout [1] was a service that measured the influence of a person using information from multiple social networks. Their initial framework [56] used long lasting (e.g., in-degree, pagerank centrality, recommendations etc) and dynamic features (reactions to a post such as retweets, up-votes etc.) to estimate the influence of a person across nine different social networks.

Further studies have tried to classify users based on factors such as celebrity status, socioeconomic status etc. Lampos et al. [39] classify the socioeconomic status of users on Twitter as high, middle or lower socioeconomic, using features such as tweet content, topics of discussion, interaction behaviour, and user impact. They obtain an accuracy of 75% using a nonlinear, generative learning approach with a composite Gaussian Process kernel. Preoctiuc-Pietro et al. [54] present a Gaussian Process regression model, which predicts the income of the user on Twitter. They examined factors that help characterize user income on Twitter and analyze their relation with emotions, sentiments, perceived psycho-demographics, and language used in posts. Further, Marwick et al. [45] qualitatively study the behaviours of celebrities on Twitter and how it impacts creation and sharing of content online. They aim to conceptualize “celebrity as a practice” in terms of personal information revelation, language usage, interactions, and affiliation with followers, among other things. There are also other studies that try to characterize usage patterns [2] and personalities [62] of varied users on Twitter.

Multiple existing studies attempt to detect and analyze automated activity on Twitter [15, 16, 20, 25, 71, 78] and differentiate bot activity from human or partial-human activity. Conversely, Chu et al. [16] identify users on Twitter that generate automated content. The verification badge was a key feature used for the purpose. Holistically characterizing features that resemble automated activity, and the extent to which exhibiting the same can hurt a user’s case for verification is further explored in Section 4.2.

Past studies on verified accounts have focused on elucidating their behaviors and properties on Twitter. Hentschel et al. [34] analyze verified users on Twitter and further use this information to identify trustworthy “regular” (not fake or spam) Twitter users. Castillo et al. [11] attempt to identify credible tweets based on a variety of profile features including whether the user was verified by the platform or not. Along similar lines, Morris et al. [51] examined factors that influence profile credibility perceptions on Twitter. They found that possessing an authenticated status is one of the most robust predictors of positive credibility. Paul et al. [52] performed multiple network analyses of the verified accounts present on Twitter and reveal how they diverge from earlier results on the network as a whole. Hence, to summarize, there exists a rich body of literature establishing the enhancement of credibility and perceived importance a verified badge endows a user with. However, no prior work, to the best of our knowledge, has attempted to characterize attributes that make the aforementioned status more attainable.

3 DATASET

In this section, we present details of our dataset and the data collection process along with a summary of the diverse features.

3.1 User Metadata

The ‘@verified’ handle on Twitter follows all accounts on the platform that are currently verified. We queried this handle on the 18th of July 2018 and extracted the IDs of 297,776 users who were verified at the time. In the interest of verifying Twitter’s assertion that likeliness of a handle’s verification is commensurate with public interest in that handle and nothing else [66, 67], we sought to obtain a random controlled subset of non-verified users on the platform. Pursuant to this need, we leveraged Twitter’s Firehose API – a near real-time stream of public tweets and accompanying author metadata – in order to acquire a random set of 284,312 non-verified users, controlling for a conventional measure of public interest, by ensuring that the number of followers of every non-verified user obtained was within 2% that of a unique verified user that we had previously acquired.

Twitter provides a REST Application Programming Interface (API) with various endpoints that make data retrieval from the site in an organized manner easier. We used the REST API to acquire profile metadata of the user handles obtained previously including account age, number of friends, followers and tweets. Additionally, we obtained the number of public Twitter lists a user was part of and the handle’s profile description. Metadata features extracted from user profiles have previously been used for classifying users and inferring activity patterns on Twitter [48, 76]. We further focused our work to the subset of users who had English listed as their profile language thus enabling us to focus on the largest linguistic demographic on the platform [50] and leaving us with 231,235 English verified users and 175,930 non-verified users.

3.2 Content Features

Utilizing Twitter’s Firehose API, we acquired all tweets authored by the aforementioned users over a one year collection period spanning from 1st June 2017 to 31st May 2018. In total, our collection process acquired roughly 494,452,786 tweets. The tweet texts were retained and any accompanying media such as GIFs were deemed surplus to requirements and discarded.

From the text we extracted linguistic and stylistic features such as the number and proportion of *Part-Of-Speech* (POS) tags, effectively obtaining a user’s breakdown of natural language component usage. Work demonstrating the importance of content features in location inference [44], tweet classification [7], and network characterization [42] further led us to extract the frequency of hashtags, retweets, mentions and external links used by each user. Prompted by studies showing that the deceptiveness of tweets could be inferred from the length of sentences constituting them [4], we computed additional features including average words per sentence, average words per tweet, character level entropy and frequency and proportion of long words (word length greater than six letters) per user.

In the interest of better discerning the emotions conveyed by the tweets authored by a user and responses they may evoke in the

User Metadata	Number of followers Number of friends Number of statuses Number of public list memberships Account age	Temporal Features	Average number of followers last year Average number of friends last year Average number of statuses last year Proportion of followers gained in last 3 months Proportion of friends gained in last 3 months Proportion of statuses generated in last 3 months Proportion of followers gained in last 1 month Proportion of friends gained in last 1 month Proportion of statuses generated in last 1 month Average duration between statuses
Content Features	Number of POS tags ¹ Frequency of POS tags ¹ Average number of words per sentence Average number of words per tweet Character level entropy Proportion of long words ² Positive sentiment score ³ Negative sentiment score ³ Neutral sentiment score ³ Compound sentiment score ³ Frequency of hashtags Frequency of retweets Frequency of mentions Frequency of external links posted	Miscellaneous Features	LIWC analytic summary score LIWC authentic summary score LIWC clout summary score LIWC tone summary score Botometer complete automation probability Botometer network score Botometer content score Botometer temporal score Tweet topic distribution ⁴

Table 1: List of features extracted per user by our framework.

¹ Part Of Speech (POS) tags include nouns, personal pronouns, impersonal pronouns, adjectives, adverbs, verbs, auxiliary verbs, prepositions and articles.

² Long words are defined as words longer than 6 letters.

³ Sentiment scores are weighted over all tweets of a user by tweet length.

⁴ Scores over 100 topics are extracted from the tweets.

potential audience, sentiment analysis presented itself as an effective tool. Sentiment gleaned from Twitter conversations has been used to predict financial outcomes [8], electoral outcomes [3] as well as the ease of content dissemination [22]. We used Vader [26], a popular social media sentiment analysis lexicon, which has previously been widely used in a plethora of applications ranging from predicting elections [3, 55] to forecasting cryptocurrency market fluctuations [59]. We extracted positive, negative and neutral sentiment scores and an additional fourth compound score, which is a nonlinear normalized sum of valence computed based on established heuristics [74] and a sentiment lexicon. All four scores are computed per user, weighted by tweet length.

3.3 Temporal Features

Existing research suggests that temporal features relating to content generation and activity levels on Twitter can be used to infer emergent trending topics [12] as well as influential users [41].

Leveraging the Twitter Firehose, we gathered fine-grained time series of user statistics including number of friends, followers and statuses, thus permitting us to compute their averages over our one year collection period. Furthermore, positing that a user’s likelihood of verification may be predicated on how ascendant their reach in

the platform is, we compute the proportion of friends and followers gained over the last one month and the last three months of our collection period. Additionally, similar trajectory encoding features are computed for tweet activity levels over the aforementioned one and three month windows, and the average time between statuses is extracted using the status count time series on a per user basis.

3.4 Miscellaneous Features

Attempting to capture qualitative cognitive and emotional cues from a user’s tweets, we acquired the four LIWC 2015 [53] summary statistics named Analytic, Clout, Authentic and Tone for each user in our dataset. The summary dimensions indicate the presence of logical and hierarchical thinking patterns, confidence and leadership, personal cues and emotional tone, respectively, in the tweets of a user. LIWC categories have been scientifically validated to perform well in determining affect on Twitter [18, 70] and have been previously used to detect sarcasm [27] and for mental health diagnoses from Twitter conversations [31].

Furthermore, positing that accounts perceived as being completely or partially automated may have a harder time getting verified, we leveraged Botometer – a flagship bot detection solution [69] that exposes a free public API. The system is trained on thousands

of instances of social bots and the creators report AUC ROC scores between 0.89 and 0.95. Botometer utilizes features spanning the gamut from network attributes to temporal activity patterns. Additionally, it queries Twitter to extract 300 recent tweets and publicly available account metadata, and feeds these features to an ensemble of machine learning classifiers, which produce a Complete Automation Probability (CAP) score, which we acquire for every user in our dataset. We also augment our dataset with the temporal, network and content category automation scores for each user.

Finally, we also look to glean into the topics that users tweet about. Topic modelling has been effectively used in categorizing trending topics on Twitter [79] and inferring author attributes from tweet content [47]. To this end, we ran the Gibbs sampling based Mallet implementation of Latent Dirichlet Allocation (LDA) [46] setting the number of topics to 100 with 1000 iterations of sampling. Although, such a topic model could be applied on a per tweet basis and subsequently aggregated by user, we find this approach to not work very well as most tweets are simply a sentence long. To overcome this difficulty, we follow the workaround adopted by previous studies by aggregating all the tweets of a user into a single document [35, 75]. In effect, this treatment can be regarded as an application of the author-topic model [60] to tweets, where each document has a single author.

3.5 Rectifying Class Imbalance

Focusing our analysis on the Twitter Anglosphere left us with a substantially skewed class distribution of 231,235 verified users and 175,930 non-verified users in our dataset. In keeping with existing research on imbalanced learning on Twitter data [30, 49], we used a two-pronged approach to rectify this – a minority over-sampling technique named ADASYN [32] which generates samples based on the feature space of the minority examples and a hybrid over and under-sampling technique called SMOTETomek which additionally also eliminates samples of the over-represented class [36] and has been found to give exemplary results on imbalanced datasets[6]. Augmenting our classifier’s training data in the aforementioned manner allowed us to attain near-perfect classification scores.

The data collected is classified and summarized in Table 1. We intend to anonymize and make this dataset accessible to the public in a manner compliant with Twitter terms, once this work is published.

4 INFERRING VERIFIED STATUS

We commence our analysis by eliminating all features that could be deemed surfeit to requirements. To this end, we employed an all-relevant feature selection model [37] which classifies features into three categories: confirmed, tentative and rejected. We only retain features that the model is able to confirm over 100 iterations.

To evaluate the effectiveness of our framework in discerning verification status of users, we examine five classification performance metrics – precision, recall, F1-score, accuracy and area under ROC curve – for five classifiers. The first two methods intended at establishing baselines were a Logistic Regressor and a Support Vector Classifier. Further, three methods were used to gauge how far the classification performance could be pushed using the features we collected. These were (1) a Generalized Additive Model

trained by nested iterations, setting all terms to smooth, (2) a Multi Layered Perceptron with 3 hidden layers of 100, 30 and 10 neurons respectively, using Adam as an optimiser and ReLU as activation and (3) state-of-the-art Gradient Boosting tool named XGBoost with a maximum tree depth of 6 and a learning rate of 0.2. The results obtained are detailed in Table 2. The first batch of results are obtained by training on the original unadulterated training split. Even without rectifying class distribution biases, we are able to attain a high classification accuracy of 98.9% on our most competitive classifier.

The second and third batches are trained on data rectified for class imbalance using the adaptive synthetic over-sampling method (ADASYN) and a hybrid over and under-sampling method (SMOTE-Tomek), respectively. The ADASYN algorithm generates samples based on the feature space of the minority class data points and is a powerful method that has seen success across many domains [33] in neutralizing the deleterious effects of class imbalance. The SMOTE-Tomek algorithm combines the above over-sampling strategy with an under-sampling method called Tomek link removal [64] to remove any bias introduced by over-sampling. This rectification did improve results, generally improving the performance of our two baseline choices and especially helping us inch closer to perfect performance with gradient boosting. However, particularly surprising was the detrimental effect of class re-balancing on the MLP classifier which in all likelihood also learned the non-salient patterns in the re-balanced data. Also unexpectedly, the ADASYN re-balancing outperformed the more sophisticated SMOTETomek re-balancing in pushing the performance limits of the support vector (89.1% accuracy) and gradient boosting (99.1% accuracy) approaches. This might be owing to the fact that the Tomek link removal method omits informative samples close to the classification boundary thus affecting the learned support vectors and decision tree splits.

Our results suggest that near perfect classification of the Twitter user verification status is possible without resorting to complex deep-learning pipelines that sacrifice interpretability.

4.1 Feature Importance Analysis

To compare the usefulness of various categories of features, we trained gradient boosting classifier, our most competitive model, using each category of features alone. While we achieved the best performance with user metadata features, content features were not far behind. Evaluated on multiple randomized train-test splits of our dataset, user metadata and content features were both able to consistently surpass 0.88 AUC. Additionally, temporal features alone are able to consistently attain an AUC of over 0.79.

The individual feature importances were determined using the Gini impurity reduction metric output by the gradient boosting model trained on the unmodified dataset. To rank the most important features reliably, the model was trained 100 times with varying combinations of hyperparameters (column sub-sampling, data sub-sampling and tree child weight) and the features determined to be the most important were noted. The most reliably discriminative features and their normalized density distributions over the values they attain are detailed in Figure 1. These features generally exhibit intuitive patterns of separation based on which an informed

Dataset	Classifier	Precision	Recall	F1-Score	Accuracy	ROC AUC Score
Original imbalanced data	Logistic Regression	0.86	0.86	0.86	0.859	0.854
	Support Vector Classifier	0.89	0.89	0.89	0.887	0.883
	Generalized Additive Model ¹	0.97	0.98	0.98	0.975	0.976
	3-Hidden layer NN (100,30,10) ReLU+Adam	0.98	0.98	0.98	0.983	0.977
	XGBoost Classifier	0.99	0.99	0.99	0.989	0.990
ADASYN class rebalancing	Logistic Regression	0.86	0.86	0.86	0.856	0.858
	Support Vector Classifier	0.89	0.89	0.89	0.891	0.891
	Generalized Additive Model ¹	0.97	0.97	0.97	0.974	0.973
	3-Hidden layer NN (100,30,10) ReLU+Adam	0.96	0.96	0.96	0.959	0.957
	XGBoost Classifier	0.99	0.99	0.99	0.991	0.991
SMOTETomek class rebalancing	Logistic Regression	0.86	0.86	0.86	0.860	0.856
	Support Vector Classifier	0.90	0.90	0.90	0.903	0.901
	Generalized Additive Model ¹	0.98	0.97	0.98	0.974	0.974
	3-Hidden layer NN (100,30,10) ReLU+Adam	0.97	0.97	0.97	0.966	0.968
	XGBoost Classifier	0.99	0.99	0.99	0.990	0.991

Table 2: Summary of classification performance of various approaches using metadata, temporal and contextual features on the original and balanced datasets.

¹ The generalized additive models were trained using all smooth terms.

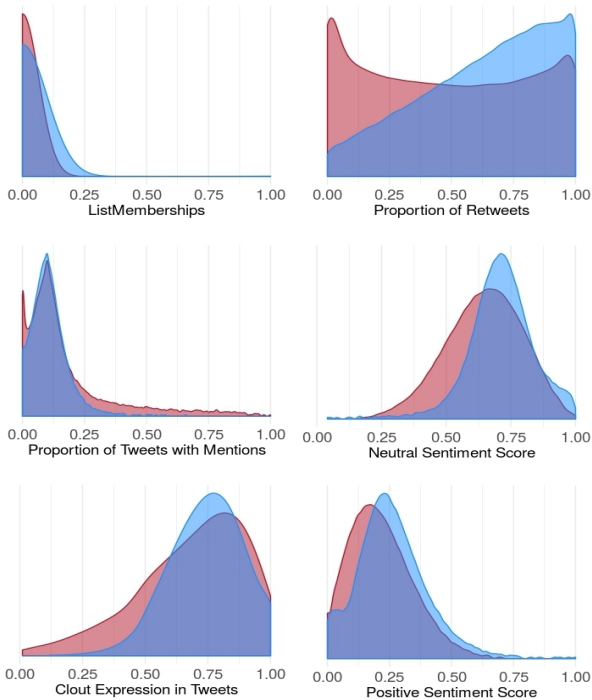


Figure 1: Normalized density estimations of the six most discriminative features for verified (blue) and non-verified users (red).

prediction can be attempted, e.g., the very highest echelons of public list membership counts are populated exclusively by verified

users while the very low extremes of propensity for authoritative speech as indicated by LIWC Clout summary scores are exclusively displayed by non-verified users.

The top 6 features are sufficient to reach performance of 0.9 AUC on their own right and the top 10 features are sufficient to further push those numbers up to 0.93. This is largely owing to the fact that substantial redundancy was observed among sets of highly correlated features such as some linguistic (tendency to use long words and impersonal pronouns highly correlate with high analytic LIWC summary scores) and temporal trajectory (most ascendant users score highly in both the 1 month and 3 month features in terms of tweets authored and followers gained) features.

4.2 Clustering and characterization

In order to characterize accounts with a higher resolution than a binary verification status will permit, we apply K-Means++ on the normalized user vectors selecting the 30 most discriminative features indicated by the XGBoost model – our most competitive classifier. We settle on 8 different clusters based on evaluation including the inflection point of the clustering inertia curve and the proportion of variance explained. In the interest of an intuitive visualization, two dimensional embeddings obtained using t-SNE dimensionality reduction method [43] are presented. Tuning the perplexity metric appropriately, the method considers the similarity of data points in our feature space and embeds them in a manner that reflects their proximity in the feature space. The embeddings are plotted and our classifier responses for members of the different clusters are detailed in Figure 2.

Investigating these clusters allows us to further unravel combinations of attributes that strengthen a user’s case for verification. Clusters C0 and C2 are composed nearly exclusively of non-verified

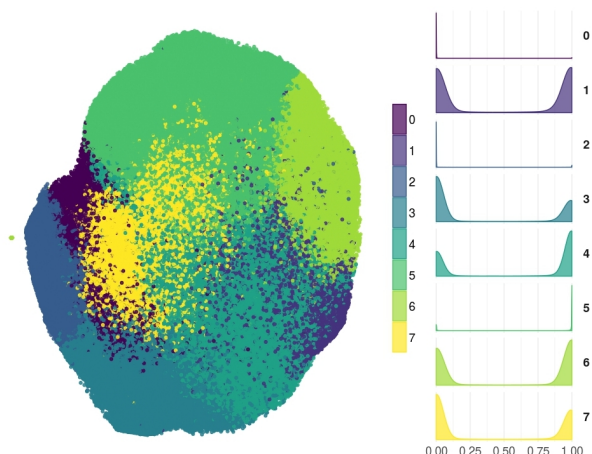


Figure 2: t-SNE embeddings of accounts coloured by cluster. The distribution of verification probabilities by cluster, as predicted by our classifier, are faceted on the right.

users. Cluster C0 can largely be characterized as the Twitter layman with a high proportion of experiential tweets. This narrative further plays out in our collected features with members of this cluster on average having short tweets, high incidence of verb usage and scoring especially high in the LIWC Authenticity summary. Cluster C2 can be characterized as an amalgamation of accounts exhibiting bot-like behavior. Members of this cluster scored highly on the complete, network and content automation scores in our feature set. Furthermore, members in C2 possessed attributes previously linked to spammers such as copious usage of hashtags [77] and external links [72]. Manual inspection verified the substantial presence of automated content such as local weather updates in this cluster. Unsurprisingly, members of this cluster were predicted to possess the lowest verification probability by our classifier.

The composition of clusters C4 and C6 leans towards verified users, with members of C4 having a tendency to post longer tweets and retweet more frequently than author content, while members of C6 almost exclusively retweet on the platform with slightly over 93% of their content being such. Cluster C5 is nearly entirely comprised of verified users and includes elite Twitteratti that comprise the core of verified users on the platform. These users have by far the highest list memberships on average while also scoring very highly on the LIWC Clout summary. Predictably, members of this cluster were predicted to possess the highest verification probability by our classifier.

The remaining clusters C1, C3 and C7 are comprised of a mix of verified and non-verified users. However, further inspection revealed that they have very divergent trajectories. Members of cluster C1 are ascendant both in terms of reach and activity levels as evidenced by the proportion of their followers gained and statuses authored in the last one and three months of our collection period. These members can be said to constitute a nouveau-elite group of users. This is further backed up by the fact that these users are lacking in their presence in public lists as compared to the very established elite in cluster C5. Manual inspection also verifies that many of these users have attained verification during our collection

Cluster	Population	Accuracy	ROC AUC Score
C0	19462	0.996	0.989
C1	26259	0.986	0.986
C2	19356	0.994	0.984
C3	46178	0.988	0.987
C4	90843	0.989	0.987
C5	105701	0.993	0.986
C6	39248	0.990	0.989
C7	60118	0.987	0.986

Table 3: Classification performance of our most competitive model broken down by cluster.

period. This is in stark contrast with members of C3 and C7 who are either stagnant or declining in their reach and activity levels and show very low engagement with the rest of the platform in terms of retweets and mentions. Remarkably, our classifier is able to make this distinction and rates members of C1 as slightly better candidates for verification on average than members of C3 or C7. The relative difficulty of classifying users in these mixed clusters is demonstrated in the performance breakdown detailed in Table 3.

5 COMPARATIVE TOPICAL USAGE ANALYSIS

Having deduced important predictive features present in a user’s metadata, linguistic style and activity levels over time with respect to verification status, we next investigate the presence of similar predictive patterns in the choice and variety of tweet topic usage amongst users.

5.1 Content Topics

In order to obtain a topical breakdown of a user’s tweets in an unsupervised manner, we ran the Gibbs sampling based Mallet implementation of Latent Dirichlet Allocation (LDA) [46] with 1000 iterations of sampling. Narrowing down on the correct number of topics T required us to execute multiple runs of the model while varying our choices for the number of topics. The model was executed for 30, 50, 100, 150 and 300 topics and the likelihood estimates were noted. It must be mentioned that in all cases the likelihood estimates stabilized well within the 1000 iteration limit we set. The likelihood keeps rising in value up to $T = 100$ topics, after which it sees a decline. This kind of profile is often seen when varying the hyperparameter of a statistical model, with the optimal model being rich enough to fit the information available in the data, yet not complex enough to begin fitting noise. This led us to conclude that the tweets we collected over a year are best accounted for by incorporating 100 separate topics. We set document-topic density $\alpha = T/50$ and topic-word density $\beta = 0.01$, which are the default settings recommended in prior studies [28] and maintain the sum of the Dirichlet hyperparameters, which can be interpreted as the number of virtual samples contributing to the smoothing of the topic distribution, as constant. The chosen value of β is small enough to permit a fine-grained breakdown of tweet topics covering various conversational areas.

Classifier	Precision	Recall	F1-Score	Accuracy	ROC AUC Score
Generalized Additive Model (GAM) ¹	0.83	0.83	0.83	0.832	0.831
3-Hidden layer NN (100,30,10) ReLU+Adam	0.88	0.88	0.88	0.882	0.880
XGBoost Classifier	0.82	0.82	0.82	0.824	0.823

Table 4: Summary of classification performance of various approaches on inferred topics.

¹ The generalized additive models were trained using all smooth terms.

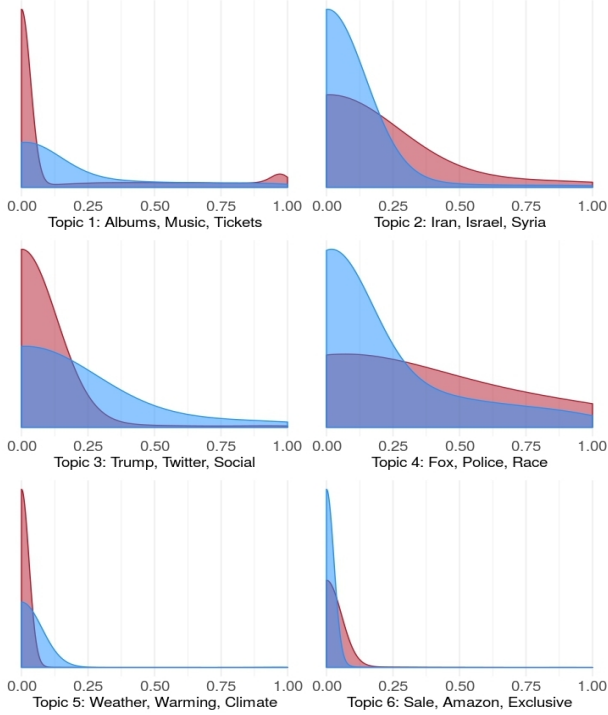


Figure 3: Normalized density estimations of usage for the six most discriminative topics for verified (blue) and non-verified users (red). Listed alongside are the top three most probable keywords for each topic.

We again commenced the prediction by pruning down our topical feature set using the all relevant feature selection method we used earlier [37] in Section 4. This allowed us to hone in on the 76 topics that were confirmed to be predictive of verification status. To evaluate the effectiveness of our framework in discerning verification status of users from topic cues, we examine five classification performance metrics – precision, recall, F1-score, accuracy and area under ROC curve – for the three classifiers that were most competitive in our previous classification task. These were (1) a Generalized Additive Model trained by nested iterations, setting all terms to smooth, (2) a Multi Layered Perceptron with 3 hidden layers of 100, 30 and 10 neurons respectively, using Adam as an optimiser and ReLU as activation and (3) Gradient Boosting tool named XGBoost with a maximum tree depth of 5 and a learning rate of 0.3. The results obtained are detailed in Table 4. The results demonstrate

that it is eminently possible to infer the verification status of a user purely using the distribution of topics they tweet about with a high accuracy. The MLP classifier was the most competitive in this task, reliably pushing past 88.2% accuracy.

In the interest of interpretability, we evaluate the predictive power of each topic with respect to the classification target. To this end, we obtain individual topic importances using the ANOVA F-Scores output by GAM – our second most competitive model on this task. In order to rank the features reliably, the procedure is run on 50 random train-test splits of the dataset and the topics with the lowest F-Scores noted. The most reliably discriminative topics and the normalized density distributions of their usage are detailed in Figure 3. Owing to multiple topics largely belonging to popular broad conversational categories such as sports and politics, some redundancy was observed in the way of multi-collinearity. This is further backed up by the fact that the top 15 most important topics alone can discern verification status with an AUC of 0.76 while the top 25 topics can push those numbers up to an AUC of 0.8 nearly approximating the GAM performance on the whole feature set (AUC 0.83). These topics generally exhibit intuitive patterns of separation based on which an informed prediction can be made, e.g., the users who tweet most frequently about climate change are all verified while controversial topics like middle-east geopolitics are something verified users prefer to devote limited attention to.

5.2 Topical Span

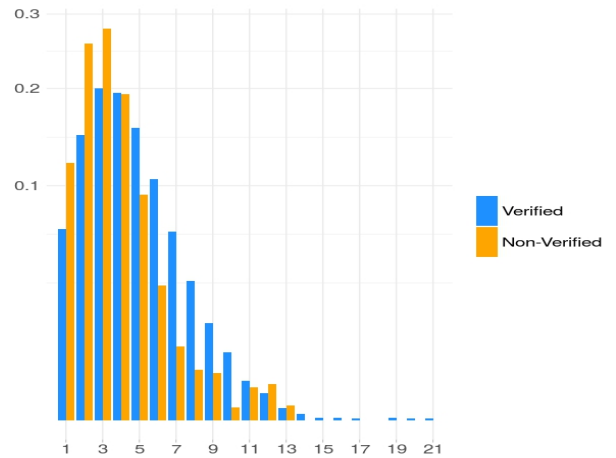


Figure 4: Square-root scaled proportion of users by optimal number of topics.

Peripheral aspects of topics such as their geographical distribution [57] and the viability of embeddings they induce for sentiment analysis [57] tasks have been explored before. This prompted us to extend our inquiry into peripheral measures such as inconsistencies in the variety and number of topics the two classes of users tweet about. In order to obtain an optimal mix of the number of topics per user in an unsupervised manner, we leveraged the use of an Hierarchical Dirichlet Process (HDP) model implementation [73] for topic inference. This method streams our corpus of tweets and performs an online Variational Bayes estimation to converge at an optimal number of topics T , for each user. Once again, we set $\alpha = T/50$ and $\beta = 0.01$, which are the default settings recommended in existing studies [28].

The distribution of cardinality for topic sets by verification status are detailed in Figure 4. Inspection of the distribution uncovers a clear trend with non-verified users clearly being over-represented in the lower reaches of the distribution (1–4 topics), while a comparatively substantial portion of verified users are situated in the middle of the distribution (5–10 topics). Also noteworthy is the fact that the very upper echelons of topical variety in tweets are occupied solely by verified users. We posit that this may be owing to the fact that news handles (e.g., ‘@BBC’: 13 topics) and content aggregators (e.g., ‘@GIFs’: 21 topics) are over represented in the set of verified users. The validation of this assertion is left for future work.

6 CONCLUSION

The coveted nature of platform verification on Twitter has led to the proliferation of verification scams and accusations of systemic bias against certain ideological demographics. Our work attempts to uncover actionable intelligence on the inner workings of the verification system, effectively formulating a checklist of profile attributes a user can work to improve upon to render verification more attainable.

This article presents a framework that computes the strength of a user’s case for verification of Twitter. We introduce our machine learning system that extracts a multitude of features per user, belonging to different classes: user metadata, tweet content, temporal signatures, expressed sentiment, automation probabilities and preferred topics. We also categorize the users in our dataset into intuitive clusters and detail the reasons behind their likely divergent outcomes from the verification procedure. Additionally, we demonstrate role, that a user’s choices and variety over conversational topics plays in precluding or effecting verification.

Our framework represents the first of its kind attempt at discerning and characterizing verification worthy users on Twitter and is able to attain a near perfect classification performance of 99.1% AUC. We believe this framework will empower the average Twitter user to significantly enhance the quality and reach of their online presence without resorting to prohibitively priced social media management solutions.

7 ACKNOWLEDGMENTS

We thank Language Technologies Research Centre (LTRC) lab at IIIT-Hyderabad and the Precog lab at IIIT-Delhi for their support.

Additionally, we would like to thank Microsoft India for granting access to their commercial Twitter solutions.

REFERENCES

- [1] 2019. Klout. <https://www.lithium.com/products/klout>. Accessed: 2019-02-16.
- [2] Hasan Al Maruf, Nagib Meshkat, Mohammed Eunos Ali, and Jalal Mahmud. 2015. Human behaviour in different social medias: A case study of Twitter and Disqus. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 270–273.
- [3] David Anuta, Josh Churchin, and Jiebo Luo. 2017. Election bias: Comparing polls and twitter in the 2016 us election. *arXiv preprint arXiv:1701.06232* (2017).
- [4] Darren Scott Appling, Erica J Briscoe, and Clayton J Hutto. 2015. Discriminative models for predicting deception strategies. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 947–952.
- [5] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 65–74.
- [6] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter* 6, 1 (2004), 20–29.
- [7] Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool, and Sungyoung Lee. 2013. Precise tweet classification and sentiment analysis. In *Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on*. IEEE, 461–466.
- [8] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.
- [9] Bustle. 2018. This Twitter Verification Scam Was Promoted By Twitter Itself, And The Consequences Are Terrifying. <https://www.bustle.com/p/this-twitter-verification-scam-was-promoted-by-twitter-itself-the-consequences-are-terrifying-7833920>. Accessed: 2018-12-27.
- [10] Claudia Canali and Riccardo Lancellotti. 2012. A quantitative methodology based on component analysis to identify key users in social networks. *International Journal of Social Network Mining* 1, 1 (2012), 27–50.
- [11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.
- [12] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining*. ACM, 4.
- [13] Meeyoung Cha, Hamed Haddadi, Fabrizio Benevenuto, and Krishna P Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *fourth international AAAI conference on weblogs and social media*.
- [14] Shelly Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology* 39, 5 (1980), 752.
- [15] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. Identifying correlated bots in twitter. In *International Conference on Social Informatics*. Springer, 14–21.
- [16] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing* 9, 6 (2012), 811–824.
- [17] Scott Counts and Kristie Fisher. 2011. Taking It All In? Visual Attention in Microblog Consumption. *ICWSM* 11 (2011), 97–104.
- [18] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.
- [19] Nicholas Diakopoulos and Arkaitz Zubiaga. 2014. Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance. In *ICWSM*.
- [20] John P Dickerson, Vadim Kagan, and VS Subrahmanian. 2014. Using sentiment to detect bots on twitter: Are humans more opinionated than bots?. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 620–627.
- [21] B Zafer Erdogan. 1999. Celebrity endorsement: A literature review. *Journal of marketing management* 15, 4 (1999), 291–314.
- [22] Emilio Ferrara and Zeyao Yang. 2015. Measuring emotional contagion in social media. *PloS one* 10, 11 (2015), e0142390.
- [23] Andrew J Flanagan and Miriam J Metzger. 2007. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society* 9, 2 (2007), 319–342.
- [24] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. ACM, 1–15.
- [25] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. 2017. An in-depth characterisation of Bots and Humans on Twitter. *arXiv preprint arXiv:1704.01508* (2017).
- [26] CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) <http://comp.social.gatech.edu/papers/icwsml4.vader.hutto.pdf>.

- [27] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, 581–586.
- [28] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1 (2004), 5228–5235.
- [29] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.
- [30] Hussam Hamdan, Patrice Bellot, and Frederic Bechet. 2015. Isislif: Feature extraction and label weighting for sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 568–573.
- [31] GACCT Harman and Mark H Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *ICWSM 2014*.
- [32] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence)*. IEEE International Joint Conference on. IEEE, 1322–1328.
- [33] Haibo He and Edwardo A Garcia. 2008. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering* 9 (2008), 1263–1284.
- [34] Martin Hentschel, Omar Alonso, Scott Counts, and Vasileios Kandydas. 2014. Finding users we trust: Scaling up verified Twitter users using their communication patterns. In *Eighth International AAI Conference on Weblogs and Social Media*.
- [35] Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. ACM, 80–88.
- [36] Miroslav Kubat, Stan Matwin, et al. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, Vol. 97. Nashville, USA, 179–186.
- [37] Miron B Kursa, Witold R Rudnicki, et al. 2010. Feature selection with the Boruta package. *J Stat Softw* 36, 11 (2010), 1–13.
- [38] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. AcM, 591–600.
- [39] Vasileios Lamos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. 2016. Inferring the socioeconomic status of social media users based on behaviour and language. In *European Conference on Information Retrieval*. Springer, 689–695.
- [40] Vasileios Lamos, Nikolaos Aletras, Daniel Preoțiu-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 405–413.
- [41] Changhyun Lee, Haewoon Kwak, Hosung Park, and Sue Moon. 2010. Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th international conference on World wide web*. ACM, 1137–1138.
- [42] Jure Leskovec and Julian J Mcauley. 2012. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*. 539–547.
- [43] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
- [44] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users. *ICWSM 12* (2012), 511–514.
- [45] Alice Marwick and Danah Boyd. 2011. To see and be seen: Celebrity practice on Twitter. *Convergence* 17, 2 (2011), 139–158.
- [46] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. (2002).
- [47] Caitlin McCollister, Bo Luo, and Shu Huang. 2015. Building Topic Models to Predict Author Attributes from Twitter Messages. In *CLEF*.
- [48] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM 11*, 5th (2011), 25.
- [49] Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. 2014. TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 628–632.
- [50] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The twitter of babel: Mapping world languages through microblogging platforms. *PLoS one* 8, 4 (2013), e61981.
- [51] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 441–450.
- [52] Indraneil Paul, Abhinav Khattar, Ponnurangam Kumaraguru, Manish Gupta, and Shaan Chopra. 2018. Elites Tweet? Characterizing the Twitter Verified User Network. *arXiv preprint arXiv:1812.09710* (2018).
- [53] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
- [54] Daniel Preoțiu-Pietro, Svitlana Volkova, Vasileios Lamos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PLoS one* 10, 9 (2015), e0138717.
- [55] Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh. 2016. Election result prediction using Twitter sentiment analysis. In *Inventive Computation Technologies (ICICT), International Conference on*, Vol. 1. IEEE, 1–5.
- [56] Adithya Rao, Nemanja Spasojevic, Zhisheng Li, and Trevor Dsouza. 2015. Klout score: Measuring influence across multiple social networks. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2282–2289.
- [57] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings. In *Thirtieth AAAI conference on artificial intelligence*.
- [58] Statista. 2018. Most popular tweets on Twitter as of November 2018, by number of retweets. <https://www.statista.com/statistics/699462/twitter-most-retweeted-posts-all-time/>. Accessed: 2018-12-22.
- [59] Evita Stenqvist and Jacob Lönnö. 2017. Predicting Bitcoin price fluctuation with Twitter sentiment analysis.
- [60] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 306–315.
- [61] S Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* 73100 (2008).
- [62] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2018. Personality Predictions Based on User Behavior on the Facebook Social Media Platform. *IEEE Access* 6 (2018), 61959–61969.
- [63] TechAcute. 2018. Top 40 List of the Most-Liked Tweets on Twitter. <https://techacute.com/list-most-liked-tweets/>. Accessed: 2018-12-22.
- [64] Ivan Tomek. 1976. Two modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics* 6 (1976), 769–772.
- [65] TripWire. 2019. Get Verified Through a Promoted Tweet? Nope. It’s a Scam! <https://www.tripwire.com/state-of-security/latest-security-news/get-verified-promoted-tweet-nope-scam/>. Accessed: 2019-1-29.
- [66] Twitter. 2018. Verified account FAQs. <https://help.twitter.com/en/managing-your-account/twitter-verified-accounts>. Accessed: 2018-12-22.
- [67] Twitter. 2019. About Verified Accounts: Twitter Help 2018. <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts>. Accessed: 2018-12-22.
- [68] Twitter. 2019. Twitter Support Statement. <https://twitter.com/TwitterSupport/status/930926124892168192>. Accessed: 2019-1-22.
- [69] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817>
- [70] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 647–653.
- [71] Alex Hai Wang. 2010. Detecting spam bots in online social networking sites: a machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 335–342.
- [72] Alex Hai Wang. 2010. Don’t follow me: Spam detection in twitter. In *2010 international conference on security and cryptography (SECRYPT)*. IEEE, 1–10.
- [73] Chong Wang, John Paisley, and David Blei. 2011. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 752–760.
- [74] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.
- [75] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 261–270.
- [76] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 705–714.
- [77] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. 2010. Detecting spam in a twitter network. *First Monday* 15, 1 (2010).
- [78] Chao Michael Zhang and Vern Paxson. 2011. Detecting and analyzing automated activity on twitter. In *International Conference on Passive and Active Network Measurement*. Springer, 102–111.
- [79] Arkaitz Zubiaga, Damiano Spina, Victor Fresno, and Raquel Martínez. 2011. Classifying trending topics: a typology of conversation triggers on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2461–2464.