The Trumpiest Trump? Identifying a Subject's Most Characteristic Tweets

Charuta Pethe
Department of Computer Science,
Stony Brook University, NY, USA
cpethe@cs.stonybrook.edu

Steven Skiena
Department of Computer Science,
Stony Brook University, NY, USA
skiena@cs.stonybrook.edu

Abstract

The sequence of documents produced by any given author varies in style and content, but some documents are more typical or representative of the source than others. We quantify the extent to which a given short text is characteristic of a specific person, using a dataset of tweets from fifteen celebrities. Such analysis is useful for generating excerpts of high-volume Twitter profiles, and understanding how representativeness relates to tweet popularity. We first consider the related task of binary author detection (is x the author of text T?), and report a test accuracy of 90.37% for the best of five approaches to this problem. We then use these models to compute characterization scores among all of an author's texts. A user study shows human evaluators agree with our characterization model for all 15 celebrities in our dataset, each with p-value < 0.05. We use these classifiers to show surprisingly strong correlations between characterization scores and the popularity of the associated texts. Indeed, we demonstrate a statistically significant correlation between this score and tweet popularity (likes/replies/retweets) for 13 of the 15 celebrities in our study.

1 Introduction

Social media platforms, particularly microblogging services such as Twitter, have become increasingly popular (Statista, 2019) as a means to express thoughts and opinions. Twitter users emit tweets about a wide variety of topics, which vary in the extent to which they reflect a user's personality, brand and interests. This observation motivates the question we consider here, of how to quantify the degree to which tweets are characteristic of their author?

People who are familiar with a given author appear to be able to make such judgments confidently. For example, consider the following pair of tweets written by US President Donald Trump, at the extreme sides of our characterization scores (0.9996 vs. 0.0013) for him:

> **Tweet 1:** Thank you for joining us at the Lincoln Memorial tonight- a very special evening! Together, we are going to MAKE AMERICA GREAT AGAIN!

> **Tweet 2:** "The bend in the road is not the end of the road unless you refuse to take the turn." - Anonymous

Although both these tweets are from the same account, we assert that Tweet 1 sounds more characteristic of Donald Trump than Tweet 2. We might also guess that the first is more popular than second. Indeed, Tweet 1 received 155,000 likes as opposed to only 234 for Tweet 2.

Such an author characterization score has many possible applications. With the ability to identify the most/least characteristic tweets from a person, we can generate reduced excerpts for high-volume Twitter profiles. Similarly, identifying the least characteristic tweets can highlight unusual content or suspicious activity. A run of sufficiently unrepresentative tweets might be indicative that a hacker has taken control of a user's account.

But more fundamentally, our work provides the necessary tool to study the question of how "characteristic-ness" or novelty are related to tweet popularity. Do tweets that are more characteristic of the user get more likes, replies and retweets? Is such a relationship universal, or does it depend upon the personality or domain of the author? Twitter users with a large follower base can employ our methods to understand how characteristic a new potential tweet sounds, and obtain an estimate of how popular it is likely to become.

To answer these questions, we formally define the problem of author representativeness testing, and model the task as a binary classification prob-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1653–1663,
Hong Kong, China, November 3–7, 2019.

lem. Our primary contributions in this paper include:

- **Five approaches to authorship verification:** As a proxy for the question of representativeness testing (which has no convincing source of ground truth without extensive human annotation), we consider the task of distinguishing tweets written by a given author from others they did not write. We compare five distinct computational approaches to such binary tweet classification (user vs. non-user). Our best model achieves a test accuracy of 90.37% over a dataset of 15 Twitter celebrities. We use the best performing model to compute a score (the probability of authorship), which quantifies how characteristic of the user a given tweet is.

- **Human evaluation study:** To verify that our results are in agreement with human judgment of how 'characteristic' a tweet is, we ask human evaluators which of a pair of tweets sounds more characteristic of the given celebrity. The human evaluators are in agreement with our model 70.40% of the time, significant above the 0.05 level for each of our 15 celebrities.

- **Correlation analysis for popularity:** Our characterization score exhibits strikingly high absolute correlation with popularity (likes, replies and retweets), despite the fact that tweet text is the only feature used to train the classifier which yields these scores.

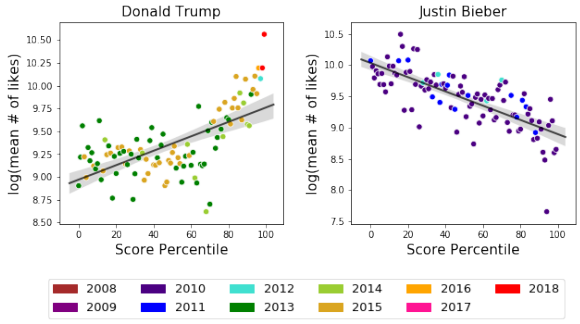

Figure 1: Plot of log mean number of likes against tweet score percentile for Donald Trump and Justin Bieber. Node color denotes the year for which the maximum number of tweets are present in each percentile bucket, demonstrating that this is not merely a temporal correlation.

For 13 of the 15 celebrities in our dataset, we observe a statistically significant correlation between characterization score and popularity. Figure 1 shows the relation between tweet score and tweet popularity for Donald Trump and Justin Bieber respectively. The figure shows that the sign of this association differs for various celebrities, reflecting whether their audience seeks novelty or reinforcement.

- **Iterative sampling for class imbalance:** Our task requires distinguishing a user's tweets (perhaps 1,000 positive training examples) from the sea of all other user's tweets (implying billions of possible negative training examples). We present an iterative sampling technique to exploit this class imbalance, which improves the test accuracy for negative examples by 2.62%.

2 Problem Formulation

We formally define the author representativeness problem as follows:

Input: A Twitter author U and the collection of their tweets, and a new tweet T.
Problem: Compute score(T, U), the probability that T was written by U. This score quantifies how characteristic of writer U, tweet T is.

2.1 Methodology

In order to obtain this representativeness score, we model our task as a classification problem, where we seek to distinguish tweets from U against tweets from all other users.

By modeling this as a binary classification problem, it becomes possible to quantify how characteristic of a writer a tweet is, as a probability implied by its distance from the decision boundary. Thus, we obtain a characterization score between 0 and 1 for each tweet.

Challenges: In training a classifier to distinguish between user and non-user tweets, we should ideally have an equal amount of examples of both classes. User tweets are simply all the tweets from that user's Twitter account, and measure perhaps in the thousands. Indeed, the number of tweets per user per day is limited to 2400 per day by current Twitter policy (https://help.twitter.com/en/rules-and-policies/twitter-limits). The negative examples consist of all tweets written by other Twitter users, a total of approximately 500 million per day

(https://business.twitter.com). Thus there is an extreme class imbalance between user and non-user tweets. Moreover, the nature of language used on Twitter does not conform to formal syntactic or semantic rules. The sentences tend to be highly unstructured, and the vocabulary is not restricted to a particular dictionary.

2.2 Data

For the binary classification task described in Section 2.1, we term tweets from U as positive examples, and tweets from other users as negative examples.

- **Positive examples:** We take tweets written by 15 celebrities from various domains, from 01-Jan-2008 to 01-Dec-2018, as positive examples. Properties of these Twitter celebrities are provided in Table 1.

- **Negative examples:** We have collected 1% of tweets from Twitter's daily feed using the Twitter API (https://developer.twitter.com/en/docs.html) to use as negative examples.

User	Tweet count (Before)	Tweet count (After)	Domain	Foll.
Amitabh Bachchan	49437	10342	Acting	37.0
Ariana Grande	37738	13657	Music	62.3
Barack Obama	11350	6772	Politics	106.0
Bill Gates	2699	1754	Business	47.1
Donald Trump	35549	18295	Politics	59.9
Ellen DeGeneres	17317	9616	TV	77.6
J K Rowling	6037	2634	Author	14.7
Jimmy Fallon	10698	3596	TV	51.1
Justin Bieber	18044	5193	Acting	105.0
Kevin Durant	22532	4146	Sports	17.5
Kim Kardashian	24541	7943	Modeling	60.5
Lady Gaga	7239	3767	Music	78.5
LeBron James	5145	2102	Sports	42.6
Narendra Modi	17613	6672	Politics	47.0
Oprah Winfrey	11685	4588	TV	42.2

Table 1: Twitter celebrities in our dataset, with tweet counts before and after filtering (Foll. denotes followers in millions)

Preprocessing and Filtering: We have preprocessed and filtered the data to remove tweets that are unrepresentative or too short for analysis. All text has been converted to lowercase, and stripped of punctuation marks and URLs. This is because our approaches are centered around word usage. However, in future models, punctuation may prove effective as a feature. Further, we restrict analysis to English language tweets containing no attached images. We select only tweets which are more than 10 words long, and contain at least 5 legitimate (dictionary) English words. We define an unedited transfer of an original tweet as a retweet, and remove these from our dataset. Since comments on retweets are written by the user themselves, we retain these in our dataset.

We note that celebrity Twitter accounts can be handled by PR agencies, in addition to the owner themselves. Because our aim is to characterize Twitter profiles as entities, we have not attempted to distinguish between user-written and agency-written tweets. However, this is an interesting direction for future research.

We use a train-test split of 70-30% on the positive examples, and generate negative training and test sets of the same sizes for each user, by randomly sampling from the large set of negative examples.

3 Related work

3.1 Author identification and verification

The challenge of author identification has a long history in NLP. PAN 2013 (Juola and Stamatatos, 2013) introduced the question: "Given a set of documents by the same author, is an additional (out-of-set) document also by that author?" The corpus is comprised of text pieces from textbooks, newspaper articles, and fiction. Submissions to PAN 2014 (Stamatatos et al., 2014) also model authorship verification as binary classification, by using non-author documents as negative examples. The best submission (Seidman, 2013) in PAN 2013 uses the General Impostors (GI) method, which is a modification of the Impostors Method (Koppel et al., 2012). The best submission (Khonji and Iraqi, 2014) in PAN 2014 presents a modification of the GI method. These methods are based on the impostors framework (Koppel and Winter, 2014).

Veenman and Li (2013) used compression distance as a document representation, for authorship verification in PAN 2013. HB et al. (2015) present a global feature extraction approach and achieve state-of-the-art accuracy for the PAN 2014 corpus. The best submission (Bagnall, 2015) in PAN 2015 (Stamatatos et al., 2015) uses a character-level RNN model for author identification, in which each author is represented as a sub-model, and the

recurrent layer is shared by all sub-models. This is useful if the number of authors is fixed, and the problem is modeled as multi-class classification. Mohsen et al. (2016) also approach multi-class author identification, using deep learning for feature extraction, and Nirkhi et al. (2016) using hierarchical clustering.

Potha and Stamatatos (2018) propose an intrinsic profile-based verification method that uses latent semantic indexing (LSI), which is effective for longer texts. Koppel and Schler (2004) and Luyckx and Daelemans (2008) explore methods for authorship verification for larger documents such as essays and novels. Nizamani and Memon (2013) and Brocardo et al. (2013) explore author identification for emails, and Chen and Sun (2017) for scientific papers. Azarbonyad et al. (2015) make use of temporal changes in word usage to identify authors of tweets and emails. Fissette (2010), Green and Sheppard (2013), and Zhang et al. (2014) evaluate the utility of various features for this task. Stamatatos (2008) proposes text sampling to address the lack of text samples of undisputed authorship, to produce a desirable distribution over classes.

Koppel et al. (2009) compare methods for variants of the authorship attribution problem. Bhargava et al. (2013) apply stylometric analysis to tweets to determine the author. López-Monroy et al. (2015) propose a document representation capturing discriminative and subprofile-specific information of terms. Rocha et al. (2016) review methods for authorship attribution for social media forensics. Peng et al. (2016a) use bit-level n-grams for determining authorship for online news. Peng et al. (2016b) apply this method to detect astroturfing on social media. Theóphilo et al. (2019) employ deep learning specifically for authorship attribution of short messages.

3.2 Predicting tweet popularity

Suh et al. (2010) leverages features such as URL, number of hashtags, number of followers and followees etc. in a generalized linear model, to predict the number of retweets. Naveed et al. (2011) extend this approach to perform content-based retweet prediction using several features including sentiments, emoticons, punctuations etc. Bandari et al. (2012) apply the same approach for regression as well as classification, to predict the number of retweets specifically for news articles. Zaman et al. (2014) present a Bayesian model for retweet prediction using early retweet times, retweets of other tweets, and the user's follower graph. Tan et al. (2014) analyze whether different wording of a tweet by the same author affects its popularity. SEISMIC (Zhao et al., 2015) and PSEISMIC (Chen and Li, 2017) are statistical methods to predict the final number of retweets. Zhang et al. (2018) approach retweet prediction as a multi-class classification problem, and present a feature-weighted model, where weights are computed using information gain.

3.3 Training with imbalanced datasets

Various methods to handle imbalanced datasets have been described by Kotsiantis et al. (2006). These include undersampling (Kotsiantis and Pintelas, 2003), oversampling, and feature selection (Zheng et al., 2004) at the data level. However, due to random undersampling, potentially useful samples can be discarded, while random oversampling poses the risk of overfitting. This problem can be handled at the algorithmic level as well: the threshold method (Weiss, 2004) produces several classifiers by varying the threshold of the classifier score. One-class classification can be performed using a divide-and-conquer approach, to iteratively build rules to cover new training instances (Cohen, 1995). Cost-sensitive learning (Domingos, 1999) uses unequal misclassification costs to address the class imbalance problem.

4 Approaches to authorship verification

As described in Section 2.1, we build classification models to distinguish between user and non-user tweets. We have explored five distinct approaches to build such models.

4.1 Approach 1: Compression

This approach is inspired from Kolmogorov complexity (Li and Vitányi, 2013), which argues that the compressibility of a text reflects the quality of the underlying model. We use the Lempel-Ziv-Welch (LZW) compression algorithm (Welch, 1984) to approximate Kolmogorov complexity by dynamically building a dictionary to encode word patterns from the training corpus. The longest occurring pattern match present in the dictionary is used to encode the text.

We hypothesize that the length of a tweet T from user U, compressed using a dictionary built

from positive examples, will be less than the length of the same tweet compressed using a dictionary built from negative examples.

We use the following setup to classify test tweets for each Twitter user in our dataset:

1. Build an encoding dictionary using positive examples ($\text{train}_{\text{pos}}$), and an encoding dictionary using negative examples ($\text{train}_{\text{neg}}$).

2. Encode the new tweet T using both these dictionaries, to obtain $T_{\text{pos}} = \text{encode}_{\text{pos}}(T)$ and $T_{\text{neg}} = \text{encode}_{\text{neg}}(T)$ respectively.

3. If the length of T_{pos} is less than that of T_{neg}, classify T as positive; else, classify it as negative.

This gives us the class label for each new tweet T. In addition, we compute the characterization score of tweet T with respect to user U, as described in Equation 1.

$$\text{score}(T, U) = 1 - \frac{\text{len}(T_{\text{pos}})}{\text{len}(T)} \quad (1)$$

Thus the shorter the length of the encoded tweet, the more characteristic of the user T is.

4.2 Approach 2: Topic modeling

We hypothesize that each user writes about topics with a particular probability distribution, and that each tweet reflects the probability distribution over these topics. We train a topic model using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) on a large corpus of tweets, and use this topic model to compute topic distributions for individual tweets. We then use these values as features. We experiment with two types of classifiers: Logistic Regression (LR), and Multi Linear Perceptron (MLP) of size $(5, 5, 5)$. We represent each tweet as a distribution over $n = 500$ topics.

The characterization score of a tweet T is given by the classifier's confidence that T belongs to the positive class.

4.3 Approach 3: n-gram probability

We hypothesize that a Twitter user can be characterized by usage of words and their frequencies in tweets, and model this using n-gram frequencies.

We use the following setup to classify test tweets for each Twitter user in our dataset:

1. Build a frequency dictionary of all n-grams in positive examples ($\text{train}_{\text{pos}}$), and a frequency dictionary of all n-grams in negative examples ($\text{train}_{\text{neg}}$).

2. Compute the average probability of all n-gram sequences in the new tweet T using both these dictionaries, to obtain $\text{prob}_{\text{pos}}(T)$ and $\text{prob}_{\text{neg}}(T)$ respectively. Here, we use add-one smoothing and conditional backoff to compute these probability values.

3. If $\text{prob}_{\text{pos}}(T)$ is greater than $\text{prob}_{\text{neg}}(T)$, classify T as positive; else, classify it as negative.

The characterization score of tweet T is given by the average n-gram probability computed using the frequency dictionary of $\text{train}_{\text{pos}}$. We experiment with $n = 1$ (unigrams) and $n = 2$ (bigrams).

4.4 Approach 4: Document embeddings

We hypothesize that if we obtain latent representations of tweets as documents, tweets from the same author will cluster together, and will be differentiable from tweets from others. To that end, we use the following setup:

1. We obtain representations of tweets as document embeddings. We experiment with two types of document embeddings: FastText (Facebook-Research, 2016) (embedding size = 100) and BERT-Base, uncased (Devlin et al., 2018) (embedding size = 768).

2. We then use these embeddings as features to train a classification model. We experiment with two types of classifiers: Logistic Regression (LR) and Multi Linear Perceptron (MLP) of size $(5, 5, 5)$.

The characterization score of tweet T is given by the classifier's confidence that T belongs to the positive class.

Iterative sampling: As described in Section 2.1, there exists an extreme class imbalance for this binary classification task, in that the number of negative examples is far more than the number of positive examples. Here, we explore an iterative sampling technique to address this problem. We train our classifier for multiple iterations, coupling the same $\text{train}_{\text{pos}}$ with a new randomly sampled $\text{train}_{\text{neg}}$ set in each iteration.

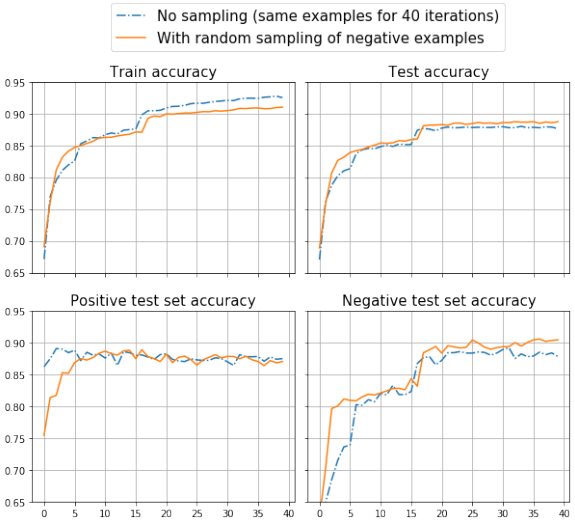

Figure 2: Mean accuracy of the BERT + MLP classifier for all users over 40 iterations

We conduct this experiment for all users with the best performing model for this approach, i.e. we use BERT embeddings as features, and MLP for classification. We train this classifier for 40 iterations, and compare the model's performance when we use the same set of negative examples vs. when we randomly sample new negative examples in each iteration.

Figure 2 shows the mean train and test accuracy for all users over 40 iterations. As expected, the training accuracy is higher if we do not sample, as the model gets trained on the same data repeatedly in each iteration. However, if we perform random sampling, the model is exposed to a larger number of negative examples, which results in a higher test accuracy (+ 1.08%), specifically for negative test examples (+ 2.62%).

4.5 Approach 5: Token embeddings and sequential modeling

In this approach, we tokenize each tweet, and obtain embeddings for each token. We then sequentially give these embeddings as input to a classifier.

We use a pretrained model (BERT-Base, Uncased: 12-layer, 768-hidden, 12-heads, 110M parameters) to generate token embeddings of size 768, and pass these to a Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) classifier. We use an LSTM layer with 768 units with dropout and recurrent dropout ratio 0.2, followed by a dense layer with sigmoid activation.

User	(1) Compression	(2) LDA + LR	(2) LDA + MLP	(3) Bigram	(3) Unigram	(4) FT + LR	(4) FT + MLP	(4) BERT + LR	(4) BERT + MLP	(5) BERT + LSTM
Amitabh Bachchan	72.93	69.47	74.91	84.45	90.16	84.58	87.13	93.59	93.73	**96.32**
Ariana Grande	71.98	76.76	80.57	73.89	85.62	84.00	85.28	87.36	87.98	**90.20**
Barack Obama	78.85	80.09	87.20	82.47	92.58	91.75	93.06	95.28	95.57	**96.57**
Bill Gates	74.29	70.78	81.78	81.21	87.10	86.34	83.97	91.56	92.41	**92.41**
Donald Trump	72.11	77.38	81.91	77.68	89.70	87.72	88.88	90.94	91.62	**93.40**
Ellen DeGeneres	70.75	69.39	74.53	71.90	84.40	81.27	83.11	87.38	88.60	**91.12**
J K Rowling	63.39	64.12	70.84	71.27	77.08	79.40	**80.75**	79.34	80.68	79.71
Jimmy Fallon	67.39	72.04	73.89	78.41	85.59	82.25	83.06	85.82	86.95	**88.62**
Justin Bieber	73.16	70.99	79.84	75.06	85.72	85.50	86.76	89.53	89.92	**92.89**
Kevin Durant	68.78	76.56	80.24	74.61	86.21	85.76	86.93	84.75	85.84	**88.62**
Kim Kardashian	69.73	72.10	76.39	71.49	83.89	80.92	82.49	84.12	85.25	**88.35**
Lady Gaga	66.01	67.07	72.40	71.44	81.14	76.91	79.90	81.46	83.21	**84.54**
LeBron James	67.95	66.21	73.33	74.17	82.42	81.97	77.05	83.48	84.77	**85.53**
Narendra Modi	82.78	84.40	89.51	90.75	94.39	94.69	95.71	97.21	**97.41**	97.33
Oprah Winfrey	68.61	61.74	70.47	75.37	83.88	83.69	83.51	86.37	87.07	**90.01**
Mean	71.25	71.94	77.86	76.95	85.99	84.45	85.17	87.88	88.73	**90.37**

Table 2: Test accuracy (%) of five approaches to classify user vs. non-user tweets (The best performing approach is shown in **bold** for each user) [Note that for each user, the test set contains an equal number of positive and negative examples.]

We train this model using the Adam optimizer (Kingma and Ba, 2014) and binary cross-entropy loss, with accuracy as the training metric.

4.6 Results and Comparison

Table 2 presents the user-wise test accuracy of the five approaches under the specified configurations. Note that the test set contains an equal number of positive and negative examples for each author.

Other baselines that we attempted to compare against include the best submissions to the PAN 2013 and 2014 author verification challenge: Seidman (2013) and Khonji and Iraqi (2014), which are variants of the Impostors Method. This challenge employed significantly longer documents (with an average of 1039, 845, and 4393 words per document for articles, essays and novels respectively, as opposed to an average of 19 words per tweet) and significantly fewer documents per author (an average of 3.2, 2.6 and 1 document/s per author, as opposed to an average of 6738 tweets per user). Our experiments with the authorship verification classifier (Eder et al., 2016) showed that the Impostors Method is prohibitively expensive on larger corpora, and also performed too inaccurately on short texts to provide a meaningful baseline.

For 13 of the 15 users in our dataset, Approach 4.5 (token embeddings followed by sequential modeling) has the highest accuracy. This model correctly identifies the author of 90.37% of all tweets in our study, and will be used to define the characterization score for our subsequent studies.

5 User study

To verify whether human evaluators are in agreement with our characterization model, we conducted a user study using MTurk (Amazon, 2005).

5.1 Setup

For each user in our dataset, we build a set of 20 tweet pairs, with one tweet each from the 50 top-scoring and bottom-scoring tweets written by the user. We ask the human evaluator to choose which tweet sounds more characteristic of the user. To validate that the MTurk worker knows enough about the Twitter user to pick a characteristic tweet, we use a qualification test containing a basic set of questions about the Twitter user. We were unable to find equal numbers of Turkers familiar with each subject, so our number of evaluators n differs according to author.

5.2 Results

Table 3 describes the results obtained in the user study: the mean and standard deviation of percentage of answers in agreement with our model, the p-value, and the number of MTurk workers who completed each task. We find that the average agreement of human evaluators with our model is 70.40% over all 15 users in our dataset.

User	Mean(%)	σ(%)	p-value	n
Amitabh Bachchan	67.08	16.44	**1.30e-07**	12
Ariana Grande	67.19	24.01	**7.60e-10**	16
Barack Obama	55.75	17.04	2.43e-02	20
Bill Gates	70.26	14.19	**1.72e-15**	19
Donald Trump	83.85	8.87	**2.52e-58**	26
Ellen DeGeneres	73.75	14.22	**5.44e-22**	20
J K Rowling	65.79	10.04	**7.51e-10**	19
Jimmy Fallon	80.00	21.93	**5.76e-25**	14
Justin Bieber	71.94	22.57	**3.97e-17**	18
Kevin Durant	64.38	15.04	**3.04e-07**	16
Kim Kardashian	71.25	14.95	**9.27e-18**	20
Lady Gaga	85.00	10.31	**1.45e-22**	9
LeBron James	63.50	12.15	**7.38e-08**	20
Narendra Modi	60.45	13.68	2.34e-03	11
Oprah Winfrey	75.79	18.12	**1.25e-24**	19

Table 3: MTurk user study results: For each of these 15 celebrities, human evaluators support our representativeness scores with a significance level above 0.05. (p-values $< 10^{-5}$ are shown in **bold**.)

For each of the 15 celebrities, the human evaluators agree with our model above a significance level of 0.05, and in 13 of 15 cases above a level of 10^{-5}. This makes clear our scores are measuring what we intend to be measuring.

6 Mapping with popularity

6.1 Correlation

We now explore the relationship between characterization score and tweet popularity for each of the users in our dataset. To analyze this relationship, we perform the following procedure for each author U:

1. Sort all tweets written by U in ascending order of characterization score.
2. Bucket the sorted tweets by percentile score (1 to 100).
3. For each bucket, calculate the mean number of likes, replies, and retweets.

4. Compute the correlation of this mean and the percentile score.

User	Likes	Replies	Retweets
Donald Trump	**0.64**	**0.63**	**0.55**
Amitabh Bachchan	**0.58**	**0.81**	**0.69**
Narendra Modi	**0.46**	0.01	0.22
Jimmy Fallon	**0.29**	**0.54**	**0.41**
J K Rowling	0.21	**0.32**	0.14
Lady Gaga	0.05	0.12	-0.01
Bill Gates	-0.05	-0.11	-0.21
LeBron James	-0.22	**-0.27**	-0.24
Oprah Winfrey	**-0.30**	**-0.41**	-0.17
Ellen DeGeneres	**-0.34**	**-0.29**	**-0.40**
Barack Obama	**-0.45**	**-0.46**	**-0.45**
Kevin Durant	**-0.57**	**-0.67**	**-0.53**
Kim Kardashian	**-0.71**	**-0.72**	**-0.70**
Justin Bieber	**-0.73**	**-0.50**	**-0.71**
Ariana Grande	**-0.74**	**-0.77**	**-0.75**

Table 4: Pearson correlation coefficients between mean popularity measure and percentile, for each user (Coefficients with p-value < 0.01 are shown in **bold** color). Green values exhibit significant positive correlation, and red values significant negative correlation.

The Pearson correlation coefficients (r-values) are listed in Table 4. The users at the top (Trump, Bachchan, Modi) all display very strong positive correlation. We name this group UPC (Users with Positive Correlation), and the group of users at the bottom (Grande, Bieber, Kardashian) as UNC (Users with Negative Correlation).

6.2 Interpretation

For users with positive correlation, the higher the tweet's characterization score, the more popular it becomes, i.e. the more likes, replies, and retweets it receives. In contrast, for users with negative correlation, the higher the tweet score, the less popular it becomes.

Figure 3 shows the plot of log mean number of likes per bucket vs. tweet score percentile, for users with the highest positive correlation. Similarly, Figure 4 shows the plot of log mean number of likes per bucket vs. tweet score percentile, for users with the highest negative correlation.

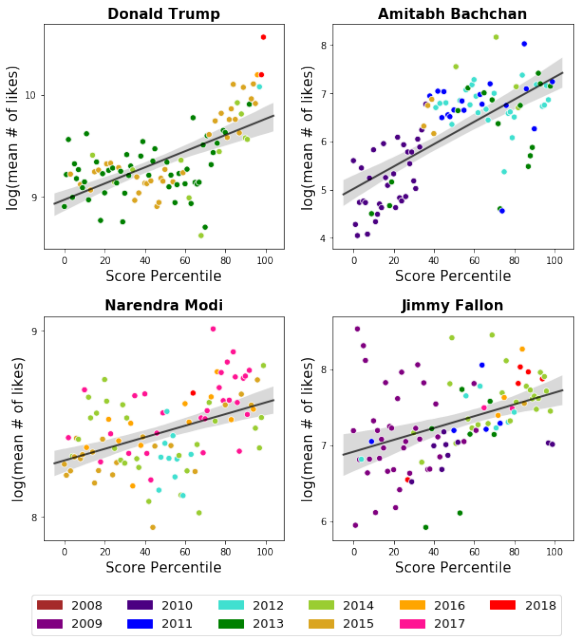

Figure 3: Log mean likes vs. percentile for users of positive correlation (The color denotes the year for which maximum tweets are present in the percentile bucket).

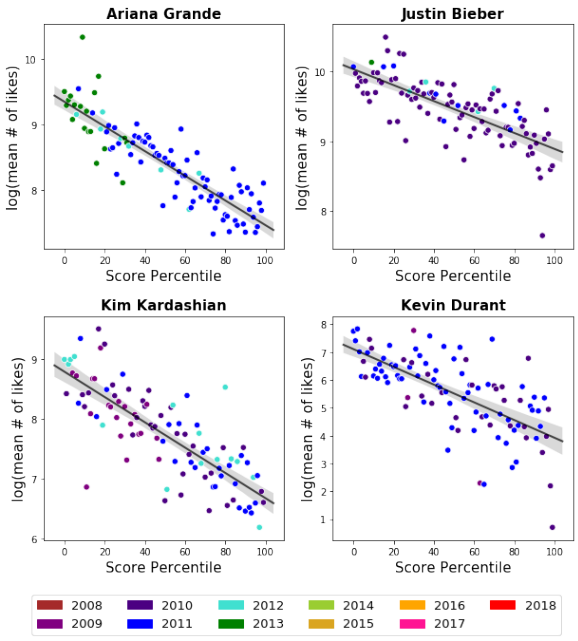

Figure 4: Log mean likes vs. percentile for users of negative correlation (The color denotes the year for which maximum tweets are present in the percentile bucket).

One may question whether these results are due to temporal effects: user's popularity vary with time, and perhaps the model's more characteristic tweets simply reflect periods of authorship. Figures 3 and 4 disprove this hypothesis.

Here the color of each point denotes the year for which most tweets are present in the corresponding bucket. Since the distribution of colors over time is not clustered, we infer that the observed result is not an artifact of temporal effects. In both cases, there is a strong trend in tweet popularity based on tweet score. We note that the plots are presented on the log scale, meaning the trends here are exponential.

6.3 Qualitative Analysis

We present examples of the most and least characteristic tweets for celebrities from three categories, along with their corresponding characterization scores computed using Approach 4.5.

6.3.1 Users with Positive Correlation (UPC)

Donald Trump

Tweet	Score
Prior to the election it was well known that I have interests in properties all over the world. Only the crooked media makes this a big deal!	0.9998
Today is the first day of the rest of your life - make the most of it!	0.0001

Amitabh Bachchan

Tweet	Score
T 2843 - The work is demanding .. the crew binding .. the city exciting .. and the dialogues expanding .. 'BADLA' is grinding .. !!	0.9996
hahaha .. now i dont have a HD .. but ya a car ride is on ..	0.0002

The characterization score appears to have correctly captured aspects of the user's personality from their corpus of tweets. For these celebrities, high scoring tweets generally prove more popular (In this example - Donald Trump: 70.5K vs. 693 likes; Amitabh Bachchan: 7.1K vs. 9 likes), as reflected in their positive correlation coefficients.

6.3.2 Users with Negative Correlation (UNC)

Ariana Grande

Tweet	Score
Finalizing the set list for Fresno! Getting so excited.. Can't believe the show is already almost sold out, you guys are amazing. Xoxo!	0.9997
The first thing I do when I get to a new city is look up how close the nearest Whole Foods is.	0.0002

Justin Bieber

Tweet	Score
grateful to everyone who came out and to my band, dancers, and whole crew. The energy last night was incredible and cant wait to tour	0.9999
Less cantaloupe, more berries. I'm talking to you, pre-packaged fruit salads. Don't play me like that.	0.00002

Again, high scoring tweets appear more characteristic of their respective users. But here, low scoring tweets are generally more popular (In this example - Ariana Grande: 622 vs. 2.4K likes; Justin Bieber: 454 vs. 13.8K likes), as reflected in their negative correlation coefficients.

6.3.3 Users with no significant correlation

Bill Gates

Tweet	Score
I recently visited a lab doing super-cool energy work-a good reminder of why governments should sponsor R&D	0.9986
There's a lot of green on this map-which is good-but still not enough.	0.0027

Here, tweets from extreme ends of the spectrum have similar content, so little variation can be expected in their popularity. For this celebrity, there is no significant correlation between characterization score and popularity.

7 Conclusions

We have presented and evaluated measures of binary author classification, to obtain a user-specific characterization score for each tweet. We demonstrate that sequential modeling on word embeddings yields the best result of 90.37% mean test accuracy, and that human evaluators are in agreement with our model 70.40% of the time. Our work demonstrates that representativeness scores correlate with popularity, and opens new research directions concerning virality on social media.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful feedback. We also thank Niranjan Balasubramanian and H. Andrew Schwartz for their comments and suggestions. This work was partially supported by NSF grants IIS-1546113 and IIS-1927227.

References

Amazon. 2005. MTurk. (https://www.mturk.com/).

Hosein Azarbonyad, Mostafa Dehghani, Maarten Marx, and Jaap Kamps. 2015. Time-aware Authorship Attribution for Short Text Streams. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 727–730. ACM.

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*.

Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The Pulse of News in Social Media: Forecasting Popularity. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. 2013. Stylometric Analysis for Authorship Attribution on Twitter. In *International Conference on Big Data Analytics*, pages 37–47. Springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6. IEEE.

Hsin-Yu Chen and Cheng-Te Li. 2017. PSEISMIC: A Personalized Self-Exciting Point Process Model for Predicting Tweet Popularity. In *2017 IEEE International Conference on Big Data*, pages 2710–2713. IEEE.

Ting Chen and Yizhou Sun. 2017. Task-Guided and Path-Augmented Heterogeneous Network Embedding for Author Identification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 295–304. ACM.

William W Cohen. 1995. Fast effective rule induction. In *Machine Learning Proceedings 1995*, pages 115–123. Elsevier.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Pedro Domingos. 1999. MetaCost: A General Method for Making Classifiers Cost-Sensitive. In *KDD*, volume 99, pages 155–164.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R: a Package for Computational Text Analysis. *R journal*, 8(1).

Facebook-Research. 2016. FastText. (https://research.fb.com/fasttext/).

MVM Fissette. 2010. Author Identification in Short Texts.

Rachel M Green and John W Sheppard. 2013. Comparing frequency-and style-based features for twitter author identification. In *FLAIRS Conference*.

Barathi Ganesh HB, U Reshma, et al. 2015. Author identification based on word distribution in word space. In *2015 ICACCI*, pages 1519–1523. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation*, 9(8):1735–1780.

Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. In *CLEF (Working Notes)*.

Mahmoud Khonji and Youssef Iraqi. 2014. A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF). *CLEF (Working Notes)*, 1180:977–983.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the Twenty-First International Conference on Machine learning*, page 62. ACM.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational Methods in Authorship Attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Yaron Winter. 2012. The "Fundamental Problem" of Authorship Attribution. *English Studies*, 93(3):284–291.

Moshe Koppel and Yaron Winter. 2014. Determining if Two Documents Are Written by the Same Author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

SB Kotsiantis and PE Pintelas. 2003. Mixture of Expert Agents for Handling Imbalanced Data Sets. *Annals of Mathematics, Computing & Teleinformatics*, 1(1):46–55.

Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36.

Ming Li and Paul Vitányi. 2013. *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media.

A Pastor López-Monroy, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Efstathios Stamatatos. 2015. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, 89:134–147.

Kim Luyckx and Walter Daelemans. 2008. Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 513–520. ACL.

Ahmed M Mohsen, Nagwa M El-Makky, and Nagia Ghanem. 2016. Author Identification using Deep Learning. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 898–903. IEEE.

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference*, page 8. ACM.

Smita Nirkhi, RV Dharaskar, and VM Thakare. 2016. Authorship Verification of Online Messages for Forensic Investigation. *Procedia Computer Science*, 78:640–645.

Sarwat Nizamani and Nasrullah Memon. 2013. CEAI: CCM-based email authorship identification model. *Egyptian Informatics Journal*, 14(3):239–249.

Jian Peng, Kim-Kwang Raymond Choo, and Helen Ashman. 2016a. Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, 70:171–182.

Jian Peng, Raymond Kim-Kwang Choo, and Helen Ashman. 2016b. Astroturfing Detection in Social Media: Using Binary n-Gram Analysis for Authorship Attribution. In *2016 IEEE Trustcom/BigDataSE/ISPA*, pages 121–128. IEEE.

Nektaria Potha and Efstathios Stamatatos. 2018. Intrinsic Author Verification Using Topic Modeling. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, page 20. ACM.

Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. 2016. Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33.

Shachar Seidman. 2013. Authorship Verification Using the Impostors Method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*. Citeseer.

Efstathios Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2):790–799.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the Author Identification Task at PAN 2015. In *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, Toulouse, France. CEUR.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the Author Identification Task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–21.

Statista. 2019. Twitter: Number of active users 2010-2018. (https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/).

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. *arXiv preprint arXiv:1405.1438*.

Antônio Theóphilo, Luís AM Pereira, and Anderson Rocha. 2019. A Needle in a Haystack? Harnessing Onomatopoeia and User-specific Stylometrics for Authorship Attribution of Micro-messages. pages 2692–2696.

Cor J Veenman and Zhenshi Li. 2013. Authorship Verification with Compression Features. In *CLEF (working notes)*.

Gary M Weiss. 2004. Mining with Rarity: a Unifying Framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19.

Terry A. Welch. 1984. Technique for High-Performance Data Compression. *Computer*, 17.

Tauhid Zaman, Emily B Fox, Eric T Bradlow, et al. 2014. A Bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8(3):1583–1611.

Chunxia Zhang, Xindong Wu, Zhendong Niu, and Wei Ding. 2014. Authorship identification from unstructured texts. *Knowledge-Based Systems*, 66:99–111.

Yang Zhang, Zhiheng Xu, and Qing Yang. 2018. Predicting Popularity of Messages in Twitter using a Feature-weighted Model. *http://ww.nlp.ia.ac.cn/2012papers/gjhy/gh154.pdf*, 20.

Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *Proceedings of the 21th ACM SIGKDD*, pages 1513–1522. ACM.

Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. 2004. Feature Selection for Text Categorization on Imbalanced Data. *ACM SIGKDD Explorations Newsletter*, 6(1):80–89.