Who's Tweeting?

Final Project Report

YAGO BARDI, CHARLES HERBERT, JOHANNA OTT, MATTEO SALVALAGGIO

ACM Reference Format:
Yago Bardi, Charles Herbert, Johanna Ott, Matteo Salvalaggio. 2020. Who's Tweeting?: Final Project Report. 1, 1 (December 2020), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ABSTRACT

This project aims to identify the authors of tweets from seven of the biggest North-American news agencies. Previous work in this field has shown that good prediction performance can be achieved with multiple author identification. The two main types of features considered in the past are style-based features, which take into account how information is communicated, and topic-based features, which take into account the content of the communication. Recent publications indicate that style-based features provide better predictive power compared to topic-based features. In this paper, topic-based features are constructed using a bag of words approach, and the style-based features are based on the features found in past research. The style- and topic-based features are combined to boost the performance of tree-based, stochastic vector machines and logistic regression models, achieving a 0.74 F1 score on unseen data. Furthermore, the models are also used to compare style-based features to topic-based features, and the results showed that better performance is obtained by using style-based features, which is consistent with the literature to date. A deeper analysis of the most relevant features for the models suggests that the links within the tweets contain the most relevant information for the classification algorithms.

Author's address: Yago Bardi, Charles Herbert, Johanna Ott, Matteo Salvalaggio.

Manuscript submitted to ACM

INTRODUCTION

Author identification is a branch of Natural Language Processing, also known as NLP, which utilizes machine learning algorithms in order to identify the author of a text based on recognizing features such as number of characters and vocabulary used.
The majority of studies in this field has concentrated on formal writings, for example books and poems. However, in the modern information age, there is an abundance of brief, colloquial writings. It has become increasingly important to establish whether identification techniques can be applied to these texts. Examples of these writings include emails, forums, blogs and social media.
There are several real world applications for author identification, for example determining the origin of anonymous messages for security purposes or using author identification as the basis for developing targeted advertising.
This project is focused on identifying the authors of tweets from US news agencies on Twitter, a social network that imposes a limit of 280 character messages per post for their users. [8] The motivation behind this is to lay a foundation for future social science and political research. By proving that the agencies can be differentiated based on their characteristics in communication and the topics they tweet about, one can expand this research for example by looking into the effects of these differences and how they affect the political division of US citizens, or more generally, how these communication styles appeal to different political views.
The analysis will be based on both the way in which these news agencies communicate, as well as the vocabulary used and topics covered by each of the agencies. Finally, this analysis will be done using supervised learning algorithms for classification.

PROBLEM DEFINITION

The problem statement of this project is as follows:

Given Twitter data
Use Natural Language Processing (NLP) techniques and classification algorithms
To to identify the author among several US news agencies

The data is collected by retrieving tweets from seven US news agencies via their main Twitter account and mainly comprises the text of the tweets. The mapping of the agency name to the Twitter account name can be found in the Table 1.

Table 1. Agency name and account name

Agency Name	Account Name
Yahoo News	YahooNews
Huffington Post	HuffPost
CNN	CNN
The New York Times	nytimes
Fox News	FoxNews
NBC News	NBCNews
Mail Online	MailOnline

There are two main constraints on the data used in this project. Firstly, Twitter restricts the number of tweets that can be retrieved from a specific user account. Secondly, since the content and topics of tweets are compared between news agencies, it is important for the tweets to come from the same time period. The findings of this project thus hold true only for time period specified in the methodology section, which is the time period for which tweets were retrievable for all the above mentioned news agencies. For more details on these constraints, refer to the methodology section of this paper.

The problem lends itself to a multi-class classification problem, which forms part of the supervised learning sphere of machine learning. The name of the news agencies are considered the classes that are to be identified. Because of the nature of the problem, the F1-Score was chosen as the measure to maximise. This is mainly because the data presents some class imbalances, further discussed in the methodology section, and the F1-Score is appropriate for measuring classification on imbalanced data.

RELATED WORK

The following section sets out the main related work we have considered for this project.

Green, Rachel M and Sheppard, John W [7] investigated whether author identification is accurate on texts of short length. It was thought that the minimum number of words needed for accurate classification was 250 words per text - much more than the average tweet length of 25 words. This paper shows that author identification is indeed possible on shorter texts. The paper further investigated the use of bag of words as a feature and compared it to using style markers. Examples of style markers include length of words, use of punctuation and length of messages. It was found that the use of style markers is more effective for author identification in tweets - not only did it result in a more accurate classification, but it was computationally much more efficient.

The paper is closely related to this project, since in both cases the final objective is author identification of tweets. Likewise, style and topic features are used to train and test the models, which is in line with the rest of this report. The major difference is the type of author it aims to predict. In this project, the focus is placed on the major North-American news agencies, while in this paper the focus on individual users related to the financial world.

Antonio Castro and Brian Lindauer [3] investigated whether the identity of a Twitter user can be predicted using linguistic stylometry. Their work is built on that from Narayanan et al. [10], and adapted to Twitter users. For this purpose, they used Twitter accounts with the same author, such that training is performed only on some of their accounts, and testing on the remaining ones. The features used focus on capturing the style of the tweet, rather than the topic, and the features importance is estimated via the Shannon Entropy. The reported results are obtained from a combination of nearest neighbours (NN) and regularized least squares classification (RLSC). The results obtained show excellent accuracy on cross-validation for the training data, and have an acceptable result on testing with unseen Twitter accounts.
A very similar approach was taken by Mishari Almishari et al. [1], who investigated the results obtained with stylised features. In their case, Naive-Bayes classifier was used.
Therefore, these papers include information on style feature selection from which to build the features for this project. While the objectives are slightly different, it is expected that a similar feature engineering will be necessary to obtain good predictions with the dataset in this project.

Apart from the above-mentioned style and topic features, there exists features that combine both of these sets, such as N-grams.[12] In general, N-gram models consist of n words and a probability assigned to the last word and the whole sequence.[11] According to Atlamimi A., Alotaibi S., and Alruban A., N-grams are a significant technique for author identification based on Twitter data.[9] N-grams were not considered for this project and are left for future work, since both topical preferences and writing style are taken into consideration for this project, and because of the time constraints.

The aim in this project is to build on the research of the author identification research by using a mixture of style markers and bag of words. The hypothesis is twofold. On the one hand it is expected that the general case as stated in the above mentioned papers - that one can predict the author based on Twitter data - also holds true for the specific case of US news agencies. The second hypothesis is that more accurate results are obtainable by using the combination of the two feature approaches compared to using only one of the approaches.

METHODOLOGY

Data set

The data set has been created by using the tweepy API to extract tweets data from the official Twitter accounts of the news agencies. Twitter only allows the last 3200 tweets to be extracted from any given Twitter handle on a free account - as a result, there were limitations on the size of the data set that could be used for this project.[14] With this in mind, the raw data was extracted twice from Twitter. As the rate of tweets sent out from agencies per day varied strongly, as can be seen in the following graphic Fig 1, the tweets for further steps were only selected from periods in which all agencies sent out tweets. This way, it was ensured that the tweets cover the same topics.

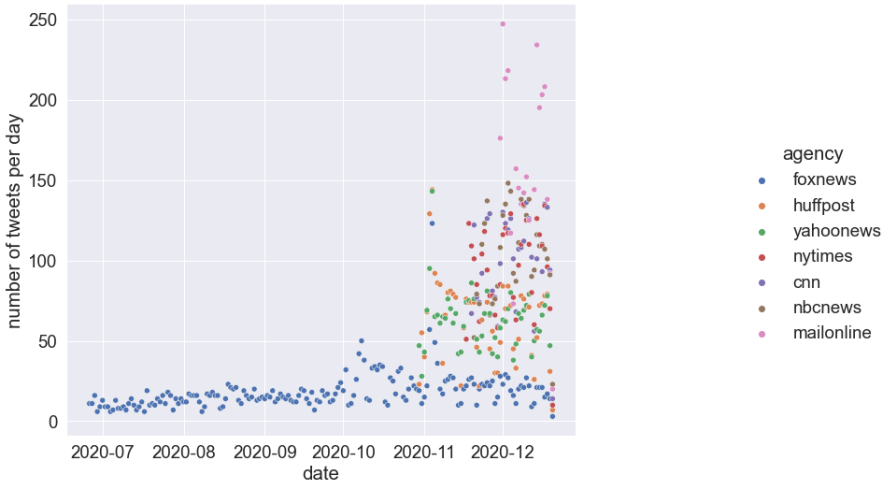

Fig. 1. Tweets per day per organisation from the second data collection

These time periods were 2020-10-12 to 2020-10-31 and 2020-11-30 to 2020-12-20.
Lastly, retweets, which is a "re-posting of a Tweet"[15], have been excluded. This has been done because retweets are not originally drafted by the agencies whose profiles the algorithms are supposed to learn, and the time constraints on this project did not allow for a further analysis of how similar the communication characteristics and topics from a retweet is to those of the agencies. Furthermore, this practice is in line with previous publications addressing Twitter author identification [5] [7]. Taking the above data decisions into account, the number of tweets in the data set amounted to 20,796 and the composition of the tweets by news agency can be seen in table 2. The class imbalance showed in this figure will be addressed in the pre-processing and feature engineering as well as in the model selection section of this paper.
The final dataset of this data collection process stores the text of the tweets (also referred to as the "tweets" from hereon) and the names of the news agencies. In addition, the metadata for every

Table 2. Relative share of the data set per agency

Agency Name	Relative Share of the Data set
CNN	21.2%
NBC News	20.7%
The New York Times	16.3%
Mail Online	14.9%
Huffington Post	11.7%
Yahoo News	11.4%
Fox News	3.8%

tweet, for instance the timestamp, amount of retweets and favourites has not been taken into account in this project.

Pre-processing and feature engineering

Two different approaches to feature selection was considered in this project, and both of these approaches had distinctly different methodologies for data pre-processing. The two approaches are the 'bag of words' approach and the 'style marker' approach, and both of these and their accompanying pre-processing and feature selection steps are discussed below.

Bag of words approach

For the bag of words approach, a large number of pre-processing steps were carried out. These steps are summarised below:

As the first step, the data was cleaned by removing links and emojis, because these were considered out of scope for this project. The tweets were then broken down into their basic units - tokens - which converts the string to an array of smaller strings. In the same step, stop-words, the most common words in a language, were removed, because they "aren't significant and distort the word frequency analysis"[13]. In the following steps, these tokens were lemmatized. This means that inflected words were replaced by their non-inflected form, for example "going" became "go". This is done so that words sharing the same base form can be grouped together, as they most likely carry a very similar meaning in a sentence. Furthermore, it serves to further reduce the dimension of the data. In a similar fashion, common synonyms that could be expected in the tweets were replaced by a single synonym. The group of synonyms can be summarized as: Donald Trump, Joe Biden, Kamala Harris and covid synonyms. For example, for covid, this included "corona","covid-19", "corona virus".

At this stage, every tweet has been mapped to a very short array of important words. This has been used to determine which words to use as features in the modelling. The next few paragraphs

outline how the final set of word features was determined.

For easing the next step, a mapping between words and integer ids was introduced by constructing a gensim dictionary [16]. Based on this, a bag of words, which is "a representation of text that describes the occurrence of words within a document"[2] was created for all agencies taking into account all their tweets. Furthermore, this enabled the creation of a list containing the words used by an agency in a tweet and the number of appearances of the words in that tweet, also known as the frequency.

For each agency, this frequency of each used word was accumulated over all tweets for that agency and then the words per agency was listed in a decreasing order of frequency. This was the basis for deciding which words to include in the feature list. Using the sorted words list, the set of words (taking into account their frequencies) that comprise 30% of the total word count per news agency was selected as features. This is not the same as taking the top 30% of the words - rather, the list is considerably smaller because there are words that appear several times. This process was repeated for every news agency. After the seven sets of word features have been determined, the sets were merged into one large set and only the unique words were kept as features. This resulted in 350 words as features. As a last note, the 30% figure was determined by testing a range of different values (30% to 60%) and selecting the point at which more features did not result in significance performance gains. This analysis is shown in the evaluation section. This greatly reduced the dimension from thousands of words to only a few hundreds of words.

Style marker approach
The pre-processing for the style features followed that reported in [3], [7] and [10]. The following were extracted from each tweet: nº of characters; nº of words; frequency of upper lower letters, as well as of upper lower words,; nº of words with-in each of the lengths between 1 and 20; frequency of each alphabetical letter and digit, as well as of some ASCII symbols symbols such as '/'; count of non ASCII-symbols. In order to have an initial idea about the importance of these features, a correlation matrix can be built. Note that there are a total of 77 style features, and therefore only those with the highest correlation with respect to the agency name are shown below.

Table 3. Top correlated features

Feature	/	Lower case words	Word length = 3	n	Words in tweet	i	e	a
Corr	0.43	-0.13	-0.12	-0.12	-0.11	-0.11	-0.11	-0.10

It can be seen that the forward slash symbol shows the highest correlation, which is not reported in [3], [5] or [10]. This suggests that links, which always contain the given symbol, have an important amount of information for predicting the user. Indeed, news agencies usually include a link to the article. By looking into the extracted data, Yahoo News usually includes two links, while the remaining agencies include usually one. This would be a possible explanation for the apparently outstanding importance of this features. In addition, Castro et al. [3], include the vowels, words in tweet and lower case words in their most important ones.

Combined pre-processing and feature engineering
The features from the bag of words approach and the style marker approach were combined to create a comprehensive feature list. No interactions between the features were considered, other than those inherent in the workings of the various classifiers. Furthermore, for the numeric features, various transformations were considered, including log transformations and square root transformations for the highly skewed variables. These transformations did not have a material impact on the results and were not considered in the final models.

Models

The models that were considered were informed by the research conducted on author classification. The models considered in this project were CatBoost [4], RandomForest Classifier, Logistic Regression, Support Vector Machines and XGBoost [6].

Regarding CatBoost, in order to tune the model, a grid-search was carried out for the parameters 'l2_leaf_reg' , 'border_count', 'score_function' and 'rsm'. In addition, the 'boosting_type' was set to 'Plain', which is optimal for large datasets. In order to ensure no over-fitting, the parameter 'random_strength', which sets the amount of randomness to use for scoring splits when the tree structure is selected, is set to 1. In addition, the depth of each iteration was kept to 6. Note that the maximum value is 16, but since CatBoost works with symmetric trees, each new level of depth increases the ramifications exponentially, increasing run-time and possibly over-fitting the model.

Next, it was decided to implement the XGBoost and SVM classifiers. The former reached a F1 score close to the best model, while the latter achieved a F1 score just under 0.70. Regarding overfitting, neither of the models (XGBoost and SVM) had this problem, since the cross validation score and F1 score are extremely close to each other. Furthermore, hyperparameter tunning was attempted for both models. However, this procedure did not result in a significant increase in the F1 score.

In order to tune the Random Forest classifier, a grid-search was carried out for the parameters 'map_depth' , 'criterion', 'class_weight', 'min_samples_leaf' and 'n_estimators'. In order to combat the class imbalance in the data, the 'class_weight' parameter was set to 'balanced_subsample', which resulted in the highest F1 score. In order to prevent over-fitting, the 'map_depth' parameter was controlled. Lastly, the over fitting was also monitored by performing a 5-fold cross validation. Overall, the Random Forest differential between the Test F1 score and the CV F1 score was relatively small, suggesting the model did not over fit.

The logistic regression model was tuned to a lesser extent than the other models. The main parameters considered for tuning were 'class weight' (for which the 'balanced' setting was found to be optimal) and 'penalty'. The penalty parameter was the main way of controlling over-fitting. Between the 'L2' and 'elasticnet' penalties, the best result was achieved by using the 'L2' penalty.

EVALUATION

The following section sets out the results of the project along with how the models were evaluated. The main metric used to evaluate the models was the F1 score. Since there was some class imbalance present, it was argued that this metric would give a more realistic view of the performance of the models.
The models were split into two parts - a training set consisting of 80% of the data, and a test set consisting of the remaining 20% of the data. A 5-fold cross validation was performed to ensure that the models do not overfit and to ensure that the models would perform well on unseen data. After training the models on the training set, the models' performance on the test set was evaluated. The main results are shown in table 4.

Table 4. F1 scores for best performing models

	CatBoost	**XGBoost**	**Random Forest**	**SVM**	**Logistic Regression**
CV mean F1 score	0.755	0.721	0.682	0.690	0.664
Test F1 score	0.740	0.730	0.701	0.680	0.677

All of the above models were trained on both the style markers and the bag of words features. The CatBoost classifier yielded the best results, closely followed by the XGBoost and Random Forest. As a further investigation of the modelling taken, the impact of only using the style marker features / bag of words features was also considered. These experiments were only conducted on two of the top performing classifiers - the CatBoost Classifier and the Random Forest Classifier. Furthermore, for the sake of brevity, cross validation scores are not considered in this section. The results are shown in table 5.

Table 5. F1 scores for different feature choices

	CatBoost	Random Forest
Style markers only	0.637	0.632
Bag of words only	0.524	0.418
Style markers and bag of words	0.740	0.701

The results are consistent with those in [7] - using style markers as features result in a more accurate classification. The results also confirm the initial hypotheses in this research project - using both style markers and bag of words features result in a stronger classifier compared to using only one of them.

Table 6. Bag of words - % of Total word count as features

	158 words 20%	**350 words 30%**	**593 words 40%**	**913 words 50%**	**1487 words 60%**
F1 score - Bag of words only	0.435	0.489	0.515	0.519	0.526
F1 score - All features	0.718	0.740	0.730	0.733	0.730

The results in table 6 show that the number of words to use as features in the bag of words approach could be reduced considerably while still maintaining a similar F1 score. When only using the bag of words feature set, there is a monotonic increase in the F1 score as the number of words increase. However, when using both the style markers and the bag of words features, a larger number of words in the bag of words did not lead to a monotonic increase in the F1 score. In fact, the optimal number of words to use was found to be 350 words, which corresponds to the words whose cumulative frequencies make up 30% of the total word count, as described in the methodology section. This discovery meant that the dimension of the features could be reduced considerably, as the first set of features used in this project comprised 1487 word features (60%). Additionally, the most important features for the CatBoost model, as given by the method *model.feature_importance_*, are depicted in Table 7.

Table 7. CatBoost model Feature Importance

Feature	/	Words in tweet	Word length = 3	Lower case words	Upper letters
Relative Importance	17.85	10.62	3.08	3.05	2.3

This shows similar results to the previously presented correlation matrix. On the same line, Antonio Castro and Brian Lindauer [3] include as well the words per tweet and frequency of lowercase

words in their top 3 of feature importance. Additionally, note that the top features are part of those considered style features. This matches with the results presented in Table 5, where higher F1 scores are achieved using the style features only as opposed to using the topic features only. This is also in line with the work presented by Green et al. [5]
Furthermore, it was decided to compare the most important feature of CatBoost with the fundamental one of XGBoost. The table below contains the most important features for XGBoost.

Table 8. XGBoost model Feature Importance

Feature	characters per tweet	nb_capitalized	nb_lower	a	t
Relative Importance	550	352	297	283	279

CONCLUSIONS

The overall best result with an F1-Score of 0.74 achieved by using CatBoost proves that the tweets can successfully be assigned to its author. Furthermore, the results in this paper show that there are multiple ways to approach the author identification challenge. It was shown that both the style markers and the bag of words approach is effective at author identification, but that the style marker approach remains the most effective between the two. Moreover, by using a combination of the two approaches, one can improve the predictive power of the classifier even further.

Future work

The work in this paper was limited to the number of tweets obtainable. An avenue for further research would be to use tweets that span a longer time horizon. Through this, one can establish whether the bag of words approach remains effective over a longer time period, since a longer time period can greatly impact the choice of features to use in the bag of words approach.
Term frequencies were used to calculate which words to use as features in the bag of words approach. Another methodology that could have been followed is the Term Frequency - Inverse Document Frequency (tf.idf) measure. This measure increases or decreases the importance of words based on the times it appears in other documents/tweets of other agencies.
In future work, additional features can be taken into account to capture even more information on which the algorithms will be trained and tested. For example, the use of emojis and a deeper analysis of links contained in tweets can be taken into consideration. Similarly, the impact of using meta-data such as the time when the tweet was sent or the number of responses could be investigated. Furthermore, and as pointed out in the related work section of this paper, features that look at writing style and topic features jointly, like N-Grams, could be introduced.

REFERENCES

[1] Mishari Almishari, Dali Kaafar, Ekin Oguz, and Gene Tsudik. Stylometric linkability of tweets. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 205–208, 2014.

[2] Jason Brownlee. A gentle introduction to the bag-of-words model. https://machinelearningmastery.com/gentle-introduction-bag-words-model/.

[3] Antonio Castro and Brian Lindauer. Author identification on twitter, 2012.

[4] Inc. Catboost. Catboost - open-source gradient boosting library. https://catboost.ai.

[5] Luke Chen, Eric Gonzalez, and Coline Nantermoz. Authorship attribution with limited text on twitter, 2017.

[6] The XGBoost Contributors. xgboost - optimized distributed gradient boosting library. http://xgboost.readthedocs.io.

[7] Rachel M Green and John W Sheppard. Comparing frequency-and style-based features for twitter author identification. In *The Twenty-Sixth International FLAIRS Conference*, 2013.

[8] Dominique Jackson. Know your limit: The ideal length of every social media post. https://sproutsocial.com/insights/social-media-character-counter/.

[9] Al Majma'ah and Saudi Arabia. Surveying the development of authorship identification of text messages. 2019.

[10] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314. IEEE, 2012.

[11] Constituency Parsing. Speech and language processing. 2009.

[12] Yunita Sari, Mark Stevenson, and Andreas Vlachos. Topic or style? exploring the most useful features for authorship attribution. In *COLING*, 2018.

[13] Taranjeet Singh. Natural language processing with spacy in python. https://realpython.com/natural-language-processing-spacy-python/.

[14] Inc. Twitter. Help with missing tweets. https://help.twitter.com/en/using-twitter/missing-tweets.

[15] Inc. Twitter. Retweet faqs. https://help.twitter.com/en/using-twitter/retweet-faqs.

[16] Radim Řehůřek. corpora.dictionary – construct word<->id mappings. https://radimrehurek.com/gensim/corpora/dictionary.html.