MSc. IT & Cognition
University of Copenhagen
Language Processing 2

Twitter Fake News Detection Using Author Profiling

Allan Misasa Nielsen <qzd426@alumni.ku.dk>
Martin Baré <mvk351@alumni.ku.dk>
Lukasz Zajaczkowski <nxv526@alumni.ku.dk>

June 9th, 2020

Abstract

Within this paper, we used a multidisciplinary approach for the task of profiling fake news spreaders from a dataset of anonymous tweets. The spreading of fake news is dangerous and can lead to radicalization of groups through isolated "echo chambers". To identify why people spread fake news, we did a psychological profile of fake news spreaders and conducted a thorough literature review to select candidate features. We implement the features we considered most important in our analysis, and achieve a consistent accuracy in fake news spreader profiling of around 0.74 using XGBoost. The most accurate features included tf-idf and frequency of clickbait. We conclude that based on textual information alone it is possible with moderate accuracy to classify fake news spreaders. Our study proves that to successfully identify fake news spreaders, network data and trust ratings should also be taken into consideration. However, it is important to consider the implications of using this type of research to filter news, as there is a fine line between the benefits and possible dangers of censoring free speech.

NB: we will denote our different sections in the paper using our initials.

Contents

1 Introduction [Joint]

Our world is more interconnected than ever, and keeping up with the news is considered a responsibility of a worldly citizen. Unfortunately, with the increasing focus on avoiding "fake news", this is becoming an increasingly difficult and time-consuming task. However, the necessity to filter and fact-check news is only one of the reasons why fake news is so problematic, and in some cases dangerous. Within this article we cover some of the reasons why fake news is dangerous, how and why it spreads,and how we can use machine learning to try and detect it.

1.1 Truth vs Belief [MB]

To identify what information is fake, it is important to distinguish between belief and truth. It has become common to use "personal belief" as a premise for argument. In philosophy this is called subjective relativism [28]. In practice it can be useful for social events, as it easily allows people to sidestep sensitive topics like religion and personal taste. However, these types of relativist arguments should not be generalized onto ethical or political issues. If anyone's personal belief can be used as a sound argument, the concept of truth becomes over-saturated and loses its societal function. Furthermore, beliefs need credible evidence backing them up to be considered truth. Just because a belief has the "possibility" to be proven true in the future, should not make it true at the time it is presented. This line of reasoning should not be considered adequate as it can be used to hold false beliefs that by chance happen to be true in the future [41]. If the argument that any unjustified false belief (a false belief without any credible evidence) can be retroactively considered truth is valid, then it is never possible to unequivocally label any belief as false. Therefore, the truth of a belief should be dependent on credible evidence at the time of presentation, and the burden of proof should be on the presenter not the receiver. In conclusion, the term Truth has a slippery definition and is at risk of becoming defined by beliefs, rather than credible facts. The argument "It is true for me" should not be considered an adequate argument in itself to back up one's beliefs in scientific, political and ethical contexts. Rather, it is important to keep the essence of truth intact through making people responsible for providing credible evidence for their arguments and not mislabelling false beliefs as truth.

1.2 Fake News Definition [MB]

When we say "fake news" in this article we will be using the definition from Dictionary.com, "false news stories, often of a sensational nature, created to be widely shared or distributed for the purpose of generating revenue, or promoting or discrediting a public figure, political movement, company, etc." [1].

1.3 How Fake News Spreads [MB]

Some researchers argue that the transition of news from written paper onto the internet leads to isolated "echo chambers' where people are sheltered from beliefs that do not match their own [4]. These echo chambers are most easily observed on Social Media sites. Social media allows people to connect and share information through their personal network and have other like-minded people confirm their prior beliefs. In fact, Social media has become a platform for people to voice their opinions and share news articles as "proof". This has shifted many people's source of news from fact-checked news sites like CNN and BBC, to articles shared by individuals in their networks. In 2016, Gottfried and Shearer found that 62% of Americans get their news from Social Media [22]. In conjunction with this, Silverman found that the majority of people believe the fake news stories they read on social media and were more likely to share fake news stories than mainstream news stories [54]. Fake news is often more salient and provocative than mainstream news, which leads it to have a farther reach of appeal and promotes clicks by scrolling social media users. Advertisements also use fake news to promote clicks, using rhetoric and fear to manipulate users to want to read more [12].

1.4 Fake News can be Dangerous [MB]

The dangers of having beliefs are internal and at worst can lead to mental turmoil and biased perception. The biggest dangers of beliefs are how they inform, direct and potentially lead to action. The problem with fake news is that, as stated earlier, the news is often charged with provocative messages and tends to be sensational in nature. After engaging with fake news, people who do not fact check their sources and integrate the false information into their belief systems can wrongly believe they have found acceptable evidence for their claims. This can lead to current problems like the anti-vaccination movement. Some anti-vaccine groups have spread information on Facebook and Twitter about the mumps, rubella and measles vaccines directly causing autism despite the medical community debunking these claims [12]. More than 40% of consumers in an American study said that the medical advice they read on social media affects how they deal with their health and these three vaccines are ones that are now often omitted from children in America [12]. This is problematic because 83% to 93% of the population need to be vaccinated for this herd protection strategy to work effectively and the people who choose to not vaccinate their kids based on falsely held beliefs are now endangering all the vaccinated kids [12].

[1] `https://www.dictionary.com/browse/fake-news`

Fake news can be used as an ideological weapon and is in many cases a form of political propaganda: "information, ideas, or rumors deliberately spread widely to help or harm a person, group, movement, institution, nation, etc"[2]. Propaganda has a long history of influencing populations and contributing to harmful actions and policies. The most infamous use of propaganda was by Hitler and his party in Nazi Germany, which facilitated the creation of a political climate that allowed for the murder of millions of people. Therefore, it is important to navigate and identify propaganda and label it accordingly.

In the Western world we value our physical possessions and often see them as extensions of ourselves. In his book, "How We Know What Isn't So," Gilovich argues that beliefs are psychologically categorized as possessions [21]. Many people consider possessions (and beliefs) to be extensions of themselves and can therefore be as protective of their beliefs as they are of their physical possessions [21]. In this way, people are not willing to relinquish their beliefs willingly through argument alone, even when presented with counterfactual evidence, in the same way that a good argument would often not be sufficient for people to give up their physical belongings.

It is important to note that freedom of belief is essential for human fulfilment. For a group or civilization to restrict the beliefs of others can be just as dangerous as them spreading fake news. However, people should be held to a higher standard and affirm their beliefs through sound arguments and verified sources. To conclude, false information advertised as truth is dangerous as it can cause wrongly justified false beliefs that can lead to harmful actions. Therefore, it is paramount that we label and categorize fake news, so citizens can make informed choices on what they choose to believe.

1.5 Psychology of Fake News Spreaders [MB]

Both Plato and Aristotle argued that one of the dangers of democracy is that demagogues, political leaders who discard rationalism to appeal to prejudices and desires of citizens, could create tyranny in modern countries by manipulating ordinary and gullible citizens [36]. This fear is now more relevant than ever, as many modern leaders manipulate information to serve their political purpose and others disregard the truth in favour of entertainment. For us to identify fake news spreaders, we first thought it necessary to identify characteristics of such spreaders. Successful fake news spreaders usually want to be perceived as authorities or experts in their relevant field [44]. They are often charismatic and present themselves as similar or relatable to the people they engage with [11][7], and aim to be perceived as trustworthy and reliable.

When a person is perceived as an expert in a given field, they are often seen as a credible source within their domain. In contrast, research shows that when there is no difference in expertise between audience and influencer, the audience is far less likely to be gullible [62]. Furthermore, some research shows that there is some leakage in trust from expertise, where an expert in one field can be as convincing in an irrelevant field to their own [48].

Perceived authority, similar to expertise, can influence the persuasiveness of fake news spreaders. Studies like the famous Milgram obedience experiment of 1974 [56], found that 26/40 participants would have shocked a person with a lethal 450V of electricity when researchers disguised as scientists told them to. Some people argue that when analyzing the experiment it is 27% not 65% who followed through blindly in the experiment [5]. This is a lower amount of people who showed blind obedience, but still an important indicator of the power of perceived authority.

Being trustworthy and reliable is an intuitive reason to believe what someone says; research shows that participants are less likely to trust two people who have provided unreliable estimates in the past [62]. The problem that fake news spreaders face is that they spread fake information to begin with, so they will likely end up in situations where they are "caught" in a lie. To remain trustworthy, some fake news spreaders will lean harder into their viewpoints and lie about past information. Some famous experiments in 1975 by Loftus, shows that false memories can be created in participants after the fact, changing their opinions [32]. Therefore, changing the memories of people or brushing over their mistakes, is a way that fake news spreaders may try to keep the trust of their followers.

Another goal of fake news spreaders is to create an "echo chamber" in a majority group, as this can be a conducive environment for spreading more fake news and convincing new people to join in their beliefs. Research supports that in one-third of trials, participants follow the consensus beliefs of groups rather than following their own perception [23]. Furthermore, it is supported that majority groups carry more weight, but smaller group's views can be more influential in some cases, and both of these instances are heightened when people are less sure of their opinions [6][58].

Research shows that to avoid lessening the effectiveness of propaganda, the rhetoric of leaders has to be properly adjusted to voice the opinion of the people [29]. Propaganda only works as a consistently effective tool for manipulation when it functions more as a microphone for existing beliefs and ideas in the population, as opposed to the creation of these beliefs in the first place [29]. In this way, propaganda is more a way to increase how extreme people's views are, rather than create new ones.

2 Related Work [LZ]

The assessment of the accuracy and credibility of textual information has been under the consideration of various disciplines, such as journalism, computer sci-

[2] https://www.dictionary.com/browse/propaganda

ence, cognitive science, and psychology for decades. [35] [34] [17]. Past approaches of detecting fake news, such as peer reviews or hiring journalistically trained assessors to filter out false information, could not keep up with the amount of information being shared on social media [8]. This combined with readers seldom regarding the veracity and credibility of information on social media, makes sites like Facebook and Twitter major platforms for the proliferation of fake news. Both the amount of data being shared and the structure of the information, renders the task of detecting fake news an area where machine learning approaches could prove beneficial. When detecting fake news, current studies differ in the features that they take into consideration. In this paper, we distinguish between three broad groups of features: textual features, features extracted from the news source (such as the reputation and trustworthiness of the publication and/or author), and network features.

Textual features can be further split into four categories: language features, lexical features, psycholinguistic features, and semantic features.

Language features include Part-of-Speech tags, n-grams or bag-of-words, all of which were bases for the earliest attempts at developing automated methods for fake news or lie detection [37][53][2][38]). Syntactical analysis, by far the most complex of all language features, has also been successfully used for deception detection by Feng et al., who achieved an 85% accuracy score when performing deep syntax analysis on a set of documents [15].

Lexical features are typically related to the vocabulary used. Common lexical statistics include the number of unique words, the number of word repetitions, average word length, the frequency of certain words in the text, tf-idf, and the number of pronouns. Depending on the source of information, additional features like the number of quotations and hashtags may be introduced to the analysis. All of these features are more advanced and require more processing power than the basic language features, but often capture stylistic information crucial when identifying authors. Lexical features have been used in multiple studies including [47] [49][43][45].

Psycho-linguistic approaches may make use of premade lexica that tag persuasive and biased language (LIWC lexicon), or capture the emotion and valence conveyed by individual words (NRC lexicon). LIWC features have been successfully used for the purpose of lie or fake news detection in studies by Hancock et al., [25] Vrij et al., [61] and Volkova et al. [59]. LIWC also captures biased language cues such as quotations, markers of certainty, subjective language, and words related to anxiety or fear. Biased language has also been studied in the context of Wikipedia articles, to identify articles that do not conform to the encyclopedic rule of neutrality [46].

Semantic features are the most abstract of all textual features. The content of a post or article can be compared against an existing knowledge base in order to calculate a similarity score between the information included in the two sources. If the content is the same, there will be a high compatibility score between the two. This method has been applied to the domain of spotting fake reviewers, by comparing the comments left by users, and information from their profiles. If there was a mismatch between the two, the comment was deemed untrustworthy. For example, this technique can be applied to find a discrepancy between a comment left by a user criticizing a hotel and their profile indicating that they have never been there. This method, utilized by Feng & Hirst [16], has achieved an accuracy score of 91%. However, this technique is limited to specific domains with enough data for semantic fact checking.

Certain semantic information can also be included in word and sentence embeddings. Embeddings can represent words or entire sentences in a high-dimensional vector space, where items that are semantically similar tend to be spatially close together. Word and sentence embeddings are not just useful for comparing information inside a document, such as spotting contradictions, but can be useful when identifying patterns across multiple texts. Because embeddings can encode information such as affect or mood, they can partially replace some of the linguistic features talked about earlier in this section. Examples of utilizing embeddings can be found in [24] and [26].

Network features and metadata constitute the second broad category of features that are used for fake news detection. They include statistics about the number of comments under posts, user engagement, usage patterns, user network analysis, etc. In [47] the authors considered features like the number of likes, shares, and comments posted within certain intervals from publication time. Because of the temporal nature of some of the metadata, network features are often used in bot detection or to monitor the spread of fake news [60]. Ma et al., have also used timestamps to track the spread of rumors on microblogging sites [33].

The urgency of identifying fake news has motivated both academics and businesses to develop multiple fake news detectors of which many are available to the public. These include SurfSafe (a browser extension that compares images in news articles), Fake News Detector AI (neural network), TrustedNews (classifies articles as, biased/untrustworthy/satire/malicious/clickbait/generated/unknown/trustworthy), Fake News Detector (open source project), Fake News Guard (combines linguistic and network features), Fake News Tracker [52] and Twitter's Fabula AI. Most of these are based on a combination of features described earlier in this section.

A large body of research focuses on the analysis of tweets exclusively. Several factors makes Twitter unique, such as the limited number of characters per tweet (280 characters max), the number of daily users, and the speed at which information proliferates. One of the earliest large-scale studies of Twitter-data was one by Castillo et al., in 2011 [9]. They analyzed tweets related to trending topics, reaching accuracy scores between 70 and 80% in detecting fake news. Kwon et al., conducted a similar study that combines temporal, structural, and linguistic features of tweets. Their careful selection of diverse features helped their algorithms achieve an accuracy score of approx. 90%, which at the time was the state of the art [30]. Those early studies have laid a solid foundation for future research that seemed to grow almost exponentially after the 2016 American presidential election. Inspired by the work of both Castillo et al., and Kwon et al., Buntain and Goldbeck applied similar methods on publicly available data to highlight that automated methods of classifying fake news can achieve a higher accuracy than professional journalists [8]. However, Shao et al., have also shown how bots can exacerbate the spreading of fake news. They have conducted a systematic study of millions of tweets to detect fake accounts that help successful fake news spreaders reach wider audiences [51]. Adopting a slightly different approach, Agrawat et al., applied algorithms developed in [57] to large scale Twitter datasets to sort users by a "reputation rank." Their approach results in only a 1% misclassification rate for real news, while also detecting the majority of fake news. Even today, Twitter continues to be the most studied social network, as exemplified by more recent papers by Nyow and Chuya [39] and Ghanem et al. [19].

3 Method

3.1 Implemented Features [Joint]

The features we used were picked through group discussion and literature reviews and were agreed upon unanimously by the group for inclusion in the final code. Within this section we will go over each feature, why we picked it, and how we implemented it. In the results section we expand upon how each feature performed individually and how they performed as a whole. The major libraries imported to run these features include Textblob, spaCy and Gensim. Other than that, we imported a few feature specific functions that are included in the description of the features below.

3.1.1 TF-IDF [LZ & AN]

[LZ]We hypothesized that one of the features that would help us distinguish between fake news and "real news" spreaders, is the language used by the author. Term Frequency - Inverse Document Frequency (tf-idf) is a lexical feature used to identify terms specific to groups or categories of documents. Because tf-idf considers more than just the count of occurrences of particular terms, it can help in finding words that are used only in specific contexts or by specific authors. For example, if multiple news spreaders (both fake and real) decide to share information on an upcoming election, looking for the word 'election' itself might not be indicative of the motives of the news spreader. On the other hand, talking more about a specific candidate might be more representative of one of the two groups. To rephrase, tf-idf highlights salient words, but only if they are not overly salient in other places.

[AN]We previously hypothesized that the frequencies of specific words could mark the differences between sources. While tf-idf embeddings are untargeted, there is still intuitive appeal in its ability to distinguish fake news. Recall that fake news tends to use repetition, appeals to specific emotions, and ad hominem, therefore there should be a statistically significant difference between the terms used subjects of the different groups which should be picked up by tf-idf.

3.1.2 POS-tagging [LZ]

Part-of-speech tagging is one of the first and most basic steps in the majority of language processing tasks. The POS-tagging process assigns a label to each word, making it possible to compare frequencies of verbs, nouns, adjectives, etc. Advanced POS taggers take word context into account, and returns additional information such as verb tenses, and intensity of adjectives. While tags disregard the semantics of tweets, they give us a glimpse into the structure of the sentences and the authors' writing styles.

3.1.3 Hashtags [MB]

Hashtags are effective in sharing information with specific audience groups and effective in sharing content with new users. Therefore, we hypothesized that there may be a difference in the amount of hashtags used by fake news spreaders and average twitter users. We also hypothesized that since hashtags are clustered by affiliation to groups, certain hashtags would be used more often by fake news spreaders than typical Twitter users. A study in 2019 found that fake news tweets in America would often contain far-right leaning ideologies, several hashtags and links to YouTube videos, all to try and unite people of similar views [13]. They also found that many of these tweets would hashtag the name of fake news spreaders like the site thegoldwaterus [13]. However, our dataset was entirely anonymous and therefore had the hashtags removed; only the count of the amount of hashtags was available for analysis. We believe that with the full hashtags present we would have had a higher accuracy in our predictions but we still decided to use the total number of hashtags by each user as a feature.

3.1.4 Emoticons [MB]

Emoticons are a modern way to show sentiment in written words. It can be difficult to portray sentiment through written word, and emoticons are a way to express human facial expressions and symbols in the English language. Since fake news spreaders are often trying to construct a "story" and distort informativeness with entertainment, we believe that they would be more likely to use emoticons to tell a more emotionally charged and better story. In this way, we hypothesized that there would be a higher average polarity of sentiment in fake news tweets as compared to regular tweets. To extract the emoticons we used the function emoji.UNICODE_EMOJI, and then used the raw emoticon images when training.

3.1.5 Clickbait [MB]

As a lot of fake news spreaders want to increase traffic to their content, we hypothesized they would favour the use of clickbait tweets. Clickbait is defined on dictionary.com as "a sensationalized headline or piece of text on the Internet designed to entice people to follow a link to an article on another web page"[3]. The purpose of headlines was originally to provide a summary of the information contained within an article. Clickbait increases engagement by using techniques such as Question-Based Headlines and Forward-reference Headlines [50]. Question-Based Headlines pose the headline as a question and Forward-reference Headlines emphasize unknown information contained in the article. In this way, clickbait is a psychologically manipulative technique to increase the amount of clicks. We used a preconfigured sklearn.svm clickbait detector trained on 10,000 clickbait articles [4]. We used the trained algorithm on our data, and imported the .txt file into our code to preserve space. We also implemented a question mark counter to try and make further predictions on possible Question-Based sentences and phrases for the purpose of trying to increase click-frequency.

3.1.6 NRC Emotion lexicon [AN]

The NRC Emotion Lexicon (EmoLex) is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done through crowd sourcing [5]. This lexicon was formatted into a dictionary. To construct a feature vector using the EmoLex, we used dictionary lookup to count the number of words in each tweet that corresponds to each of the emotions and sentiments. The output is a 1x10 vector, where each element is the count of occurrences of words that are associated with each individual emotion in each document. Statistical analyses of emotions in fake vs real news have revealed a significant difference between the prevalence of appeals toward joy, disgust and anger in text bodies [42] [18], therefore the prevalence of words associated with those emotions should serve as a discriminating feature.

3.1.7 Named entities [MB]

Named entity recognition is a system that allows for the identification and labelling of organizations, dates, times, people's names, currency, etc.[6] We believed that these would be useful features as naming people and things directly within a story makes it more engaging and entertaining for the listener [21]. Furthermore, gossip is a popular way for people to interact socially. Some studies found that as much as 68% percent of conversations overheard at a college campus are about absent people [27]. The social function of gossip is rich, and includes information sharing and entertainment [27]. To create gossip it is important to use named entities so that the news is shared across conversations and leads to further engagement with the source of information. Additionally, named entities will often be used in biased news, like politically charged smear campaigns. Therefore, we believe that different named entities may be beneficial as features for fake news detection. To extract Named entities we used the built-in function in TextBlob and extracted the numerical count as integer values.

3.1.8 User Mentions [MB]

The user mentions feature represents the amount of users that are referred to or directly contacted in the tweets. We believe that as fake news spreaders are looking to tie their tweets to charismatic characters, they are more likely to utilize user mentions to address charismatic "experts" who support their views. Therefore, we hypothesized that user mentions could be a powerful indicator of associations between groups of pre-established fake news spreaders. However, since we had anonymous data, we could only extract the total number of user mentions, not the specific users that are mentioned.

3.1.9 Tweet length[Joint]

This represents the length in characters of the tweet. We have noticed that Trump often hits the upper tweet limit and thereafter uses several tweets to get his message across. Therefore, we wanted to assess whether tweet length would increase the accuracy of our prediction as a feature.

3.1.10 Unique words [AN]

We counted the number of unique words in each document to create the vocabulary size of an user. Recall that repeating a narrative, and repeating emotional

[3] https://www.dictionary.com/browse/clickbait?s=t

[4] https://github.com/nitin-cherian/Webapps/blob/master/Codementor.io/GarethDwyer/apps/clickbait/clickbait_classifier.ipynb

[5] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[6] Named entities include: ORG, GPE, MONEY, PERSON, DATE, NORP, FAC, LOC, PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, TIME, PERCENT, QUANTITY, ORDINAL, and CARDINAL.

appeals are common traits among fake news spreaders, therefore we posit that vocabulary size could capture such patterns in conjunction with tf-idf embeddings.

3.1.11 Exclamation marks and capital letters [MB]

We counted exclamation marks as they are used textually to express excitement and in some cases show anger. We believe that the use of more exclamation marks would produce more salient text and therefore be used more often by fake news spreaders. Capital letters serve a similar function as exclamation marks, but along with expressing excitement, capitalization is also eye catching as it has thicker lines and is larger than surrounding text, drawing the attention of the eye. For example, avid fake news spreaders like Donald Trump sporadically use capital letters and accentuate their tweets with exclamation points. We believe that these techniques are likely used by other fake news spreaders and therefore extracted them as features. We counted the amount of exclamation marks as an integer, and we found the percent of the tweets that are capitalized.

3.1.12 Sentiment [MB]

Sentiment analysis is the act of trying to quantify the attitude and emotional state/expression of a text or speech. Some researchers have found that the expression of sentiment is difficult to mimic and fake [25]. Ott et al., found that people who consciously write negative deceptive opinion spam can be detected with a 86% accuracy [40]. This is because they overuse negative terminology as opposed to "real news" spreaders writing on an emotionally charged subject. We wanted to use sentiment in our analysis, as research seems to indicate that it is relatively easy to recognize and difficult for fake news spreaders to fabricate.

We used Textblob which has a built-in and pre-trained sentiment analysis. The Sentiment analysis contains both polarity and subjectivity ratings.

3.1.13 Embeddings [LZ]

The last set of features that we included in our analysis were word embeddings. We initially tried using 300-dimensional pre-trained GloVe word embeddings, hoping that they would capture the mood/sentiment of each tweet. However, they did not seem to improve our results and we had to resort to training our own embeddings on the small dataset. We used gensim to create 10-, 25-, 30-, 50-, and 100-dimensional embeddings. When used as the only set of features for classification, the 25-dimensional embeddings were the most effective. Each user was then described by a 25-dimensional tf-idf weighted average of all the word embeddings. This vector was included in our final analysis alongside other features described previously.

3.2 Classifiers [AN]

Tweets are inherently noisy which renders regression-based classifiers prone to over-fitting, partially due to the bias-variance trade-off. This led us to employ ensemble classifiers such as random forests, which also partially addressed our shortcomings of having a small training set. Ensemble learning involves the averaging of several models (in this case decision trees), which reduces the risk of choosing the wrong hypothesis, e.g. type 1 errors [14]. While SVMs, logistic regression and clustering algorithms have historically proven successful in various binary classification tasks, these algorithms fall short for the task of detecting fake news as there is no clear-cut line between fake and real news. Rhetorical devices may have the same form and follow the same patterns.

We applied a stratified 10 fold cross validation on a user-level. We tested five different classifiers, the first being a linear SVM classifier, then the stock random forest classifier from the SKLearn library, the third being a random forest classifier with hyperparameters as optimized by a randomized search over different parameterizations, the fourth AdaBoost, and lastly, XGBoost. A comparison of the performance of the different classification algorithms is shown in figure 1 in the results section.

3.3 Data pre-processing [LZ]

We have split the data cleaning process into several stages, as the extraction of different features requires text in different forms. We first removed the Xml tags surrounding every tweet, and concatenated the tweets of each individual user together. With all the tweets posted by one user in one document, we could count the total number of question marks, exclamation points, hashtags, and user mentions. This has also made it possible to extract the total length of all tweets (since we had 100 tweets from each user, we did not need to worry about averages).

Having extracted information that relies on symbols like '?', '!', or '', we removed all non-alphanumeric characters from the documents. We then calculated the percentage of capital letters relative to the total number of characters used. Having done that, we changed all the characters to lowercase, as we noticed that spaCy has trouble assigning proper POS-tags if most of the words are capitalized. Finally, we loaded the documents into spaCy to extract named entities. At a later step, we have also lemmatized the words before turning them into embedding vectors. Although not strictly necessary, lemmatization gets rid of plurals and inflections, increasing the term frequency of some words.

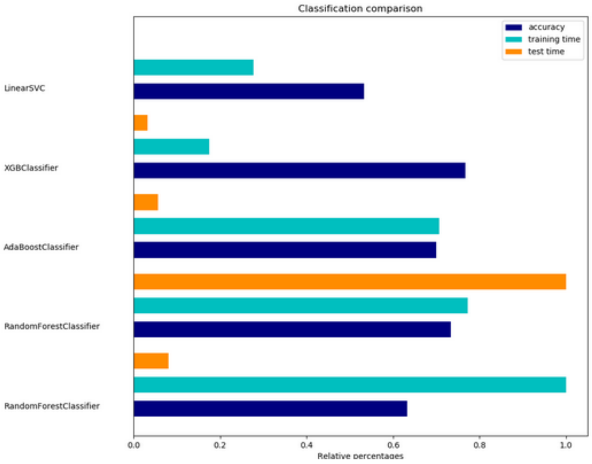


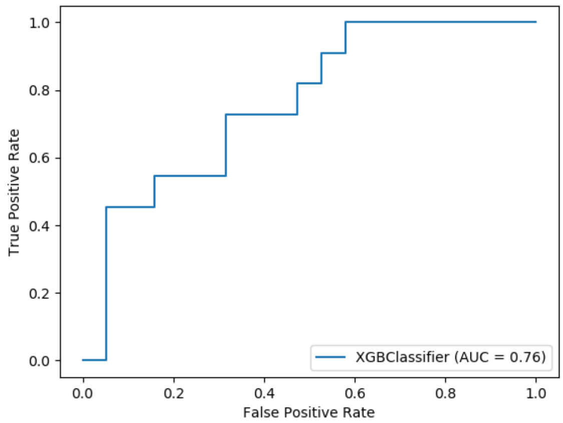
Figure 1: Classification comparisons

4 Results [AN]

Classifier	Accuracy	Training Time (in seconds)	Testing Time (in seconds)
SVM	53.3%	0.049	**< 0.001**
AdaBoost	70.0%	0.123	0.007
Random Forest (Stock)	63.3%	0.174	0.010
Random Forest (Optimized)	73.3%	0.139	0.112
XGBoost	**76.7%**	**0.030**	0.004

We have opted to show training and testing times in addition to the accuracies, as the performance of a fake news detection algorithm in a real-world scenario is judged based on its trade-off between accuracy and speed. To combat the spread of fake news, there cannot be long windows of opportunity for it to be read and shared. In figure 1, we display test and training times in relative terms to each other, and accuracies as percentages. The results are summarized in Table 1. Best results in each category are highlighted in bold. In order to secure a fair comparison, all classifications are run on the same shuffle of folds, and none of them are run with parallel processing. SVMs fail to achieve a high accuracy on the features. That is most likely due to there does not exist a clear line between fake and "real" news. The boosted tree classifiers achieve competitive results, with XGBoost achieving the best scores all-round. The intelligent penalization of trees in conjunction with the proportional shrinkage of leaf nodes may prove beneficial for our task.

Successful fake news detection algorithms should not only endeavor to maximize the accuracy score, but also keep the ratio of true to false positives as high as possible. To this end, we plot a receiver operating characteristic (ROC) curve to illustrate the diagnostic capability of the classification algorithm.



There are a lot of features used in the classification. It would be beneficial to look at the relative feature performance in terms of their contribution to the mean decrease in impurity (MDI), a metric graphed in figure 2 (appendix). The feature column names are on the x-axis, which have been reduced to indices to maintain readability. We sum up the top 10 features with highest contribution to MDI in the following table:

Feature	MDI
17 - Numerical values	0.082
58 - TF-IDF	0.064
45 - Named language	0.051
12 - Adverbs	0.045
52 - Anticipation	0.041
53 - Clickbait percent	0.040
3 - Hashtags	0.034
57 - TF-IDF	0.027
67 - TF-IDF	0.025

5 Discussion [Joint]

The relatively low scores achieved by analyzing only textual features are a testament to the complexity of detecting fake news. Purely linguistic approaches would not be robust enough at stopping the spread of misinformation, and better multidisciplinary methods must be developed. The most robust systems presented in the related works section combines virtually all kinds of features, making it more difficult for bots and fake news spreaders to elude safeguarding mechanisms.

It must be pointed out that while classification accuracy is important, it cannot be the sole metric for evaluating the strength of fake news detection systems. Researchers should also strive to develop transparent systems that minimize the number of real news that are classified as fake ones. Failing to do so would stifle free speech on social media, and the algorithms designed to act as the bulwark of a modern society against misinformation would be more detrimental to the democratic process than fake news itself.

5.1 Limitations [Joint]

A possible limitation is that we had to assume that our definition was consistent with the one used in the task. The task description did not specify what definition of "Fake News" was used when classifying the tweets and in our literature review we found many different possible definitions of the term, and would have benefited from knowing if the organizers classified things like, misleading statements, satire and hyperbole as fake news.

The limited size of the dataset poses a number of problems. One is that identifying fake news spreaders involves working with minute differences in language and emotion - types of tasks that typically require vast data sources. Another problem is that we have so little textual information that we cannot rely on long-term dependencies throughout the corpus.

Three possible methods of handling this limitation include sourcing more tweets, data augmentation and transfer learning. For instance, universal sentence encoders [10] could serve as a promising model for augmentation of data for transfer learning, so our model could learn from a variety of sources of fake news, expanding the training data beyond the confines of Twitter.

Textual data is only a small fraction of the information that is included in each tweet. Unfortunately, other descriptors that would be beneficial have been removed from the dataset, forcing us to rely on lexical and linguistic analyses only. Previous studies have successfully detected bots or fake news spreaders by considering network information (such as IP address, or the accounts followers and followees), or images. Unfortunately, the tweets excluded the actual hashtags being used, the user mentions, and the full links to the articles being shared.

The lack of timestamps proved to be especially problematic, as we could not ascertain if the tweets were listed chronologically. This completely excluded one of the more novel approaches of detecting fake news spreaders through capturing the emotional or moral narrative conveyed by a series of tweets. Research has shown that emotionally induced LSTMs can be used to identify patterns that seem to be more prevalent among fake news spreaders [19] [20]. Users who intentionally try to spread fake news tend to develop compelling and emotionally-charged stories that span multiple tweets, instead of objectively presenting information.

5.2 Future work [AN & LZ]

[AN]Crafting sentence embeddings by means of universal sentence encoders, could augment the training data with data for transfer learning, sourced from a variety of news sources outside of Twitter. The universal sentence encoder would be able to encode varied source data into universal embedding vectors that captures the semantics of sentences sourced from different authors and media, which could then be fit on a classification algorithm and thereby transfer the learning from external sources beyond the test/target source [10].

[LZ]Analyzing the narrative weaved by authors, briefly mentioned in the Limitations section, has tremendous potential in detecting fake news spreaders on twitter. Multidisciplinary research should combine theories of cognitive science, psychology, and NLP to study how people structure and present information constructed for deception. [LZ]Even though both text and images have been considered separately in the context of fake news detection, the field is currently lacking in multi-modal approaches. Frameworks such as VisualBert [31], that integrate linguistic and image features, could result not just in better algorithms for filtering news and stories, but a better understanding of what cognitive processes underlie the urge of spreading misinformation and sowing discord.

6 Conclusion [Joint]

To conclude, it is evident that fake news detection can be done with moderate accuracy on textual information alone. That said, it is important to emphasize that the research that we are conducting in identifying fake news could also be used for immoral purposes. There is a fine line between acceptable censorship and violations of basic human rights, and author profiling can be dangerous when used by governments or large organizations that are looking to coat their propaganda in a veneer of credibility. Also, heavily censoring people who spread fake news could potentially lead extremist groups to create and use even more isolated echo chambers designed specifically for the purpose of echoing their beliefs, leading to further radicalization. This would expand the already widening divide between people of different beliefs. Human freedom of expression is an integral part of being an informed citizen, and to limit it can lead to isolation of people at a personal level, and an authoritarian rule of governments at a societal level.

References

[1] R. Agrawal, L. de Alfaro, G. Ballarin, S. Moret, M. Di Pierro, E. Tacchini, and M. L. Della Vedova, "Identifying fake news from twitter sharing data: A large-scale study", *arXiv preprint arXiv:1902.07207*, 2019.

[2] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques", in *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, Springer, 2017, pp. 127–138.

[3] B. Al Asaad and M. Erascu, "A tool for fake news detection", in *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, IEEE, 2018, pp. 379–386.

[4] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election", *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[5] L. T. Benjamin Jr and J. A. Simpson, "The power of the situation: The impact of milgram's obedience studies on personality and social psychology.", *American Psychologist*, vol. 64, no. 1, p. 12, 2009.

[6] M. Biernat and M. Manis, "Shifting standards and stereotype-based judgments.", *Journal of personality and social psychology*, vol. 66, no. 1, p. 5, 1994.

[7] T. C. Brock, "Communicator-recipient similarity and decision change.", *Journal of personality and social psychology*, vol. 1, no. 6, p. 650, 1965.

[8] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular twitter threads", in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, IEEE, 2017, pp. 208–215.

[9] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter", in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[10] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, *et al.*, "Universal sentence encoder", *arXiv preprint arXiv:1803.11175*, 2018.

[11] S. Chaiken, "Heuristic versus systematic information processing and the use of source versus message cues in persuasion.", *Journal of personality and social psychology*, vol. 39, no. 5, p. 752, 1980.

[12] L. Chiou and C. Tucker, "Fake news and advertising on social media: A study of the anti-vaccination movement", National Bureau of Economic Research, Tech. Rep., 2018.

[13] M. Chong, "Discovering fake news embedded in the opposing hashtag activism networks on twitter:# gunreformnow vs.# nra", *Open Information Science*, vol. 3, no. 1, pp. 137–153, 2019.

[14] T. G. Dietterich, "Ensemble methods in machine learning", in *International workshop on multiple classifier systems*, Springer, 2000, pp. 1–15.

[15] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection", in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, Association for Computational Linguistics, 2012, pp. 171–175.

[16] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility", in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 338–346.

[17] B. J. Fogg and H. Tseng, "The elements of computer credibility", in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1999, pp. 80–87.

[18] P. Gerbaudo, "Fake news and all-too-real emotions: Surveying the social media battlefield", *Brown J. World Aff.*, vol. 25, p. 85, 2018.

[19] B. Ghanem, S. P. Ponzetto, and P. Rosso, "Factweet: Profiling fake news twitter accounts", *arXiv preprint arXiv:1910.06592*, 2019.

[20] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles", *ACM Transactions on Internet Technology (TOIT)*, vol. 20, no. 2, pp. 1–18, 2020.

[21] T. Gilovich, *How we know what isn't so*. Simon and Schuster, 2008.

[22] J. Gottfried, M. Barthel, E. Shearer, and A. Mitchell, "The 2016 presidential campaign—a news event that's hard to miss", *Pew Research Center*, vol. 4, p. 2016, 2016.

[23] R. A. Griggs, "The disappearance of independence in textbook coverage of asch's social pressure experiments", *Teaching of Psychology*, vol. 42, no. 2, pp. 137–142, 2015.

[24] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining", *Neural Computing and Applications*, pp. 1–16, 2020.

[25] J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication", *Discourse Processes*, vol. 45, no. 1, pp. 1–23, 2007.

[26] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, "A retrospective analysis of the fake news challenge stance detection task", *arXiv preprint arXiv:1806.05180*, 2018.

[27] F.-M. Hartung and B. Renner, "Social curiosity and gossip: Related but different drives of social functioning", *PLoS One*, vol. 8, no. 7, 2013.

[28] D. A. Hunter, "A practical guide to critical thinking", *Canada: John Wiley &Sons*, 2009.

[29] I. Kershaw *et al.*, *TheHitler myth': image and reality in the third Reich*. Oxford: Clarendon Press; New York: Oxford University Press, 1987.

[30] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media", in *2013 IEEE 13th International Conference on Data Mining*, IEEE, 2013, pp. 1103–1108.

[31] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visualbert: A simple and performant baseline for vision and language”, *arXiv preprint arXiv:1908.03557*, 2019.

[32] E. F. Loftus, “Leading questions and the eyewitness report”, *Cognitive psychology*, vol. 7, no. 4, pp. 560–572, 1975.

[33] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, “Detect rumors using time series of social context information on microblogging websites”, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 1751–1754.

[34] S. R. Maier, “Accuracy matters: A cross-market assessment of newspaper error and credibility”, *Journalism & Mass Communication Quarterly*, vol. 82, no. 3, pp. 533–551, 2005.

[35] A. A. Memon, A. Vrij, and R. Bull, *Psychology and law: Truthfulness, accuracy and credibility*. John Wiley & Sons, 2003.

[36] H. Mercier, “How gullible are we? a review of the evidence from psychology and social science”, *Review of General Psychology*, vol. 21, no. 2, pp. 103–122, 2017.

[37] R. Mihalcea and C. Strapparava, “The lie detector: Explorations in the automatic recognition of deceptive language”, in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Association for Computational Linguistics, 2009, pp. 309–312.

[38] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, “Lying words: Predicting deception from linguistic styles”, *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[39] N. X. Nyow and H. N. Chua, “Detecting fake news with tweets’ properties”, in *2019 IEEE Conference on Application, Information and Network Security (AINS)*, IEEE, 2019, pp. 24–29.

[40] M. Ott, C. Cardie, and J. T. Hancock, “Negative deceptive opinion spam”, in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, 2013, pp. 497–501.

[41] R. Parikh and A. Renero, “Justified true belief: Plato, gettier, and turing”, in *Philosophical explorations of the legacy of Alan Turing*, Springer, 2017, pp. 93–102.

[42] J. Paschen, “Investigating the emotional appeal of fake news using artificial intelligence and human contributions”, *Journal of Product & Brand Management*, 2019.

[43] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news”, *arXiv preprint arXiv:1708.07104*, 2017.

[44] R. E. Petty, J. T. Cacioppo, and R. Goldman, “Personal involvement as a determinant of argument-based persuasion.”, *Journal of personality and social psychology*, vol. 41, no. 5, p. 847, 1981.

[45] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, “A stylometric inquiry into hyperpartisan and fake news”, *arXiv preprint arXiv:1702.05638*, 2017.

[46] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, “Linguistic models for analyzing and detecting biased language”, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 1650–1659.

[47] J. C. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, and E. Cambria, “Supervised learning for fake news detection”, *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76–81, 2019.

[48] R. M. Ryckman, W. C. Rodda, and M. F. Sherman, “Locus of control and expertise relevance as determinants of changes in opinion about student activism”, *The Journal of Social Psychology*, vol. 88, no. 1, pp. 107–114, 1972.

[49] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, “Twitterstand: News in tweets”, in *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, 2009, pp. 42–51.

[50] J. M. Scacco and A. Muddiman, “Investigating the influence of “clickbait” news headlines”, *Engaging News Project Report*, 2016.

[51] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, “The spread of fake news by social bots”, *arXiv preprint arXiv:1707.07592*, vol. 96, p. 104, 2017.

[52] K. Shu, D. Mahudeswaran, and H. Liu, “Fakenewstracker: A tool for fake news collection, detection, and visualization”, *Computational and Mathematical Organization Theory*, vol. 25, no. 1, pp. 60–71, 2019.

[53] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective”, *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[54] C. Silverman, “This analysis shows how viral fake election news stories outperformed real news on facebook”, *BuzzFeed news*, vol. 16, 2016.

[55] S. Sriram, “An evaluation of text representation techniques for fake news detection using: Tf-idf, word embeddings, sentence embeddings with linear support vector machine.”, 2020.

[56] M. Stanley, “Obedience to authority”, *An Experimental View*, 1974.

[57] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks", *arXiv preprint arXiv:1704.07506*, 2017.

[58] E. Trouche, E. Sander, and H. Mercier, "Arguments, more than confidence, explain the good performance of reasoning groups.", *Journal of Experimental Psychology: General*, vol. 143, no. 5, p. 1958, 2014.

[59] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "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)*, 2017, pp. 647–653.

[60] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online", *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[61] A. Vrij, S. Mann, S. Kristen, and R. P. Fisher, "Cues to deception and ability to detect lies as a function of police interview styles", *Law and human behavior*, vol. 31, no. 5, pp. 499–518, 2007.

[62] I. Yaniv and E. Kleinberger, "Advice taking in decision making: Egocentric discounting and reputation formation", *Organizational behavior and human decision processes*, vol. 83, no. 2, pp. 260–281, 2000.

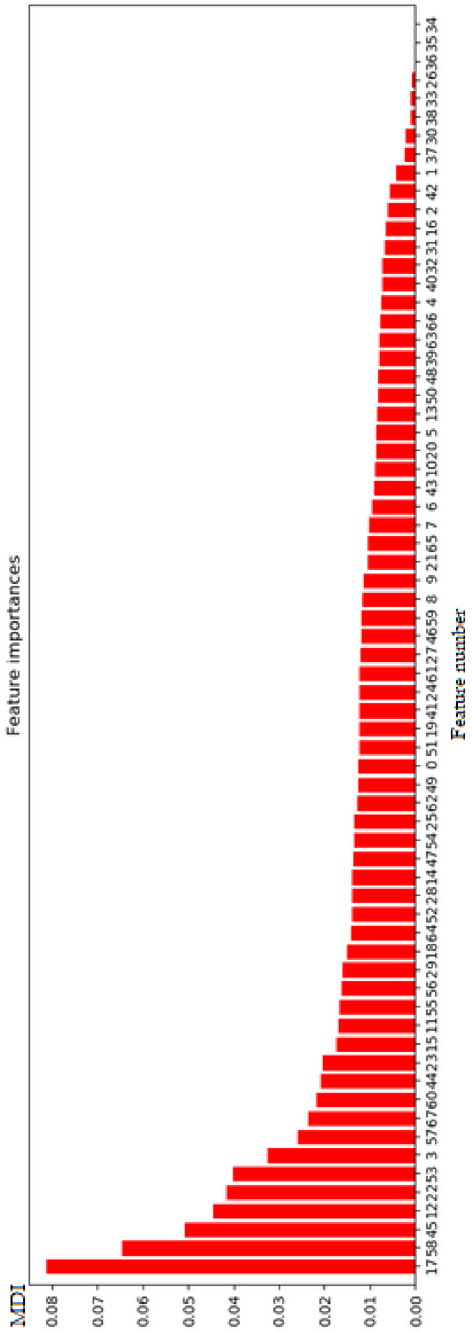


Figure 2: Feature Importances

(1) Imagine you want to spread some fake news. Imagine that all social media have installed trust indicators and once a user publishes some fake news, the user is kicked out. How would you avoid being kicked out when spreading your own fake news?

In this paper we learned a lot about the intricacies of fake news, how it spreads and how to detect it. Through this, we have learnt some ways in which we could avoid detection. A large scale study by Agrawal, et al. to classify tweets in the wild, found that pre-established association with untrustworthy news distributors was an effective way to flag fake news. When using seed lists of low quality news sites from Metacert and Opensources, the algorithm had a 90% accuracy when classifying fake news [1]. This shows that the association of news with pre-identified fake news distributors is a reliable way to identify our post as fake news. This would also be easy to identify with tags, links and follows. Therefore, to avoid getting spotted, it would be essential to avoid associating with any flagged fake news distributor. Furthermore, we would want to keep our own trust rating as high as possible. This trust rating could be established by including hyperlinks to trusted news sites; it would likely be possible to find some articles that do not contradict the narrative that we are spreading. Tf-idf in isolation was the most accurate feature in author profiling, with an accuracy score of around 0.65. Therefore, we would identify the words that are specific to fake news articles and make sure to avoid them. Our findings are supported by the widespread use of tf-idf in state-of-the-art fake news classification [3][55].

Another important feature in isolation is the clickbait classifier with an accuracy slightly higher than 0.58. From this we gather that avoiding question-based and forward-reference headlines would also aid in avoiding author profiling.

As we stated in our report, sentiment is difficult to fake. Therefore, to cheat a fake news detector, it would be beneficial to keep the sentiment of the posts as neutral as possible by trying to write as objectively as possible. We think that we could achieve this consistently by copying the headlines of credible news sites and changing a few words, such as named entities, verbs, or adjectives, to fit our narrative.

Even the most credible news sources, like BBC, have opinion pieces which would likely be flagged by fake news detectors. Therefore, detectors would require a threshold of how many instances of non-factual information can be disregarded before the author is labelled as untrustworthy. We did not include a threshold in our detection, but most filters would have to. Therefore, keeping our fake news amount under such a threshold would make it easy to write-off fake news as innocuous mistakes, and consequently would make us less vulnerable to author profiling techniques.

All the above techniques would help to avoid the detection by author profiling algorithms. The things that we believe would be most difficult to control, are the reputation of the followers we would gather, and the potential of becoming so large that we would be on the radar of more scrupulous human reviewers.

(2) We suppose that in some cases, you may not have explored some possibilities because of limitations in terms of computer power and data size. Suppose that you have unlimited power and data. What do you think could be useful in order to improve your current model?

When evaluating this question, we assume that a larger data size would imply having more data in the form that we have received it during the task, with the metadata of each tweet preserved.

As mentioned in the limitations section, we have not used algorithms that analyze data sequentially. If timestamps or other temporal information was available, we could use more advanced models like LSTMs which would be beneficial in detecting features that elude other algorithms. Neural networks also require a larger amount of data to stabilize, and a few hundred users with 100 tweets per user would not be enough for that. We found that the format of the data was not the only problem, but also the quantity of it. If we had unlimited data and processing power we would have leaned much harder in the neural network direction. Neural networks like LSTMs are the start-of-the-art in detecting time sensitive features. If we finally managed to accrue enough data for the training of a neural network to be worthwhile, a large amount of computing power would still be necessary, as it would take a long time to train (several hours to a few days).

Another direction we would have liked to pursue is using an ensemble of different classification algorithms. Each algorithm could pick up on features that other algorithms could not. For instance, discrete features could be distinguished by an SVM, and continuous features would be detected by regression techniques. Broader features could be identified by tree-based classifiers.

The data that we received included exclusively textual information without any network information, and this data was made anonymous for the purposes of the challenge. We believe that not having access to this type of data would create a glass ceiling, that we could only overcome with personal data. With unlimited computing data, it would potentially be possible to trace back the authors of the tweets, allowing us to gather further information and improve our algorithm. This would violate the purpose of the task, but it would give us information on the associations that we have previously mentioned. Most users tend to leave online trails in the form of emails, links to other websites, or posts that divulge personal information. Ethical issues aside, it would be feasible to identify the websites frequented by each user, which would enrich our dataset with additional information about each Twitter account. Establishing a connection between Twitter accounts and accounts from other social networking sites would give us extra information about the users' interests, political stance, or even what events they have attended in the past. We could theoretically evaluate the users' friend ~~networks or interest groups that they are part of to detect entire clusters of fake news spreaders that consume~~ and proliferate similar kinds of information.