Language Processing II

Group 8

Automatic Fake News Spreader Profiling

Detecting Fake News Spreaders Based on Twitter Posts

Vejleder: Manex Aguirrezabal Zabaleta & Jürgen Wedekind

Institut for Nordiske Studier & Sprogvidenskab. June 14, 2023

Group Members

Name	KUid
Jonas Kjeldmand Jensen	xgv866
Helena Björnesjö	zvw935
Tudor Laurentiu Dascalu	hpj557

Abstract

In this paper, we present a machine learning framework for detection of fake news posts on the online social media platform Twitter. We classify tweets on both an English and a Spanish dataset. While testing a variety of different methods and combinations, the best performing model for detecting fake news tweets is Gradient Boosting & Random Forrest model with features extracted from a TF-IDF approach. For English, we achieved an accuracy score of 75.33% mean accuracy and for Spanish 77.66% mean accuracy.

Division of Labour

For this project, we deem that every group member have contributed equitably. Every member has taken active part in every step and in every part of the project - from programming, to literature research, to report writing. In order to comply with the requirements, we added our names to each section.

Program Implementation

For this report, we have devised a Python code script. We have made the script publicly available to anyone interested via **github**. The script holds everything necessary for interested readers to explore the dataset for themselves.

1 Introduction (*Everyone*)

With the proliferation of online information exchange and news dissemination, the challenge of securing information validity of online news content has become an increasingly pressing matter [20, 2].

Social media is a rising source of news to many and a source of collective opinion formation, due to the growth of digital social networks. However, coinciding with this trend, big social network platforms like Twitter or Facebook have admitted to the fact that their networks are riddled with fake and cloned accounts, fake likes and other modes of non-human interactions [4].

The presence of inauthentic content on online social media platforms is especially concerning considering that these platforms are prone to fostering user confirmation bias and the emergence of *echo chambers*, social network structures forming homogeneous clusters that propagate selective, polarised content [24, 7].

Misinformation comes in many shapes and guises. Propaganda has been an important tool to authoritarian leaders and resistance movements throughout history. More recently, phenomena such as spam campaigns, hoax articles and astroturfing, have further surfaced in the digital realm. These phenomena present challenges to individual online visitors, as they are not always readily decipherable or separable from authentic, well-founded content. Tentatively, we suggest that common to all of these spurious phenomena is the intention to mislead or misinform by presenting severely biased information, information lacking trustworthy sources or evidence, or information that is knowingly false to the user behind its original posting. As such, it is apt to view massive online misinformation as a societal risk with the potential to destabilise democratic institutions, to indirectly cause harm to citizens in moments of acute crisis, to skew public opinion, and to undermine public trust in authentic news disseminated on social media platforms [9, 20, 7, 25].

Given the abundance of online social media data, access to accurate and reliable methods for automatic detection of online propaganda, fake news, or rumours without solid foundation is key to securing the validity of news outlet dissemination. Propagation of misinformation in social media can be addressed not only by identifying individual instances of misinformation (such as singular tweets or posts), but also by profiling authors, or accounts, associated with systematic spreading thereof. Authorship analysis can contribute to the identification and exclusion of malicious spreaders from social media platforms in the strides to combat large scale online distribution of misinformation.

In this paper, we direct our focus to fake news recognition. In particular, we address the challenge of distinguishing between Twitter accounts tweeting factual versus non-factual content. Twitter in particular has been subject to the spreading of misinformation [20]. We present a machine learning framework for automatic detection of fake news spreaders by attempting to identify distinctive writing style signatures of particular users. We do so by extracting lexical, sentiment and document relevant term features from user specific tweet documents that are then used for training and classification by a Random Forest model, a Gradient Booster model, and a Dense Neural Network respectively.

2 Fake News and Related Forms of Misinformation (*Jonas/Helena*)

In recent times, the concepts of *fake news* and *post-truth* have become popular catch phrases. The literature on massive online misinformation presents significant overlaps between fake news, propaganda, rumour and hoax [25, 3, 20].

In broad terms, fake news depends on the lack of credibility of a tweet or other post [4, 18]. Low credibility negatively affects information quality and may result from missing evidence or other grounding such that a proposition made in a post cannot be straightforwardly verified [24]. These properties would seem to be characteristic of rumour and fake news alike. Similarly, Hamidian & Diab [12] define rumour in terms of the spreading of content for which the truth-value cannot be determined, or else can be determined to be false. On the face of it, this holds true of fake news also. In another paper, Zubiaga et al. [29], emphasise the relevance of the disposition of the receiver of rumour messages. The difficulty of verification and the appearance of credibility of the posted information aside, they stress the effects that the message has on the receiver. For example, in a moment of crisis, such as during a hurricane, rumours might cause scepticism or anxiety in the reader. This reaction component would not seem to be necessary for fake news, albeit not in conflict with it. As with fake news, rumours can also be used to indirectly discredit or denigrate a person, group or movement [12]. Relatedly, fake news can be motivated by either the desire to smudge an opponent's reputation, or simply create confusion and deception about a topic for various reasons [16, 22].

Similarly, Volkovva et al. [25], point to another potentially distinguishing factor of fake news. The authors suggest that suspicious news is generally directed at building narratives, rather than presenting factual reports of events or states of affairs. This arguably holds true of propaganda as well.

The objective of propaganda may be viewed as the usage of communication techniques to shape information in such a way that it serves to foster an implicit predetermined agenda [14, 27], with political, religious, or ideological motivations [25]. Such agendas may or may not be associated with fake news and rumour spreading. More generally, propaganda can be seen as an attempt to manipulate the attitudes, opinions, or actions of a receiver relating to a particular topic by presenting biased messages. Alternatively, propaganda can manifest itself as an attempt to divert attention away from some information [25]. Again, these latter characteristics are also affiliated with fake news spreading.

Considering the overlap between these related phenomena of misinformation, formulating a definition of fake news such that it can be separated from rumour or propaganda is not a simple task. For the purposes of this paper, we view fake news as a form of fabricated misinformation which gives the appearance of presenting factual information. Fake news spreaders are users or bot accounts that propagate fake news through online channels, such as social media networks. For the purposes of this paper, we sidestep effects that fake news postings may have on the reader since the data we are working with does not enable us to investigate such effects.

In this project, we investigate whether, in the online Social Media context, fake news spreaders can be accurately identified and separated from authentic users with the help of Machine Learning Methods. We hypothesize that it could.

3 Related Work (*Everyone*)

In this section, we describe a selection of works on automatic misinformation detection and give a glimpse at a growing field demonstrating continued advancements.

3.1 Recent Work on Automatic Misinformation Detection (*Jonas/Helena*)

In recent years, much work has been devoted to automatic detection of spurious information being distributed on social media platforms. Topics that have been addressed range from the modeling of interaction patterns and topological structures of social networks, stance detection, user polarity, various linguistic features, to incorporating insights from cognitive psychology [28, 10, 5, 30, 11, 17].

Antoniadis et al. [3] studied misinformation spreading during emergency using Twitter data from around the time of hurricane Sandy in 2012. The authors accounted for user features like number of followers, as well as tweet features, such as number of words and number of retweets, and expected that suspicious tweets will propagate faster through a social network via retweeting patterns. Using a Bootstrap Aggregation and a Random Forest model for classification, they were able to identify misinformative tweets based on these features.

In [12], Hamidian & Diab rumour detection on Twitter data relating to Twitter posts revolving around speculations about former American president Barack Obama being Muslim was investigated. They used features including emoticons, hashtags and URLs, as well as sentiment features and part-of-speech tags. Their support vector machine (SVM) Tree Kernel Model used for classification was able to recognize rumour posts on topic relevant Twitter data.

In [25], Volkova et al. it was found that authentic and suspected news tweets can be discriminated between by modeling social interaction patterns and linguistic features relating to bias, morality and sentiment. The sentiment features were used to measure word subjectivity and opinion word polarity. Among the bias features were semantic hedges, and factual and implicative verbs, the latter two relating to the implicated truth-value of a proposition. The moral features included value pairs such as harm and care, betrayal and loyalty, and subversion and authority. Using Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for classification, they found that inauthentic news tweets contained more hedges, more high sentiment intensity words, and more moral pairs than did the verified news tweets.

In another paper, Orlov & Litvak [20] implemented an unsupervised machine learning framework for identification of Twitter users employing the platform as a conduit for spreading of propaganda. Especially significant propagandists were identified by measuring PageRank centrality in a given social network.

Del Vicario et al. [9] studied misinformation spreading on Facebook through the lens of user behavior and polarising content. The framework includes features for user behaviour such as number of users engaging with some specific Facebook post. Linguistic features such as number of words and punctuation signs, and sentence length, were also incorporated. Moreover, sentiment features such as sentiment intensity and sentiment distance between maximum and minimum sentiment intensity where also accounted for. Behavior and sentiment features in particular, contributed to the overall power of the different classifiers they implemented for topic and fake news recognition, out of which logistic regression displayed best performance.

In [18], the authors outline a logistic regression based approach for detecting whether a text is propagandist or not. They further utilise features derived from TF-IDF, BERT and various other semantic features to assist them in differentiating between the two categories.

4 The Dataset (***Tudor***)

The data used for this project consists of English and Spanish Twitter posts by a total of 600 users, split evenly between the two languages. The data includes 100 tweets with an average length of 100 characters per user. Each user is labeled as either a fake news spreader or as a non-fake news spreader. The data is balanced as the count for each of the labels is the same, 150 fake news spreaders and 150 non-fake news spreaders.

5 Methods (***Everyone***)

In this section, we present a walk-through of the methods used for extracting and selecting features based on the dataset, as well as of the models used for classification. The objective is to detect and classify fake news spreaders and being able to separate them from non-fake news spreaders by utilising features that are most informative of user identification [22]. We test our models on both English and Spanish text data, enabling us to validate model performance and further test for any possible language dependencies.

5.1 Feature Extraction (***Helena***)

In order to build a vectorised representation of the tweets relevant to this task, we employ both manual and automatic feature extraction techniques. Based on our intuition and various sources previously discussed, we decided to calculate the length of each tweet. Each tweet was then tokenised to enable computing part-of-speech (POS) tagging for structural text analysis and sentiment analysis [6]. Finally, we opted for a holistic approach and concatenated the tweets for each user in order to compute TF-IDF and Doc2Vec encoding.

Using the NLTK *SentimentIntensityAnalyzer* module, we computed text sentiment polarity as mean polarity of positive, negative and neutral sentiment valences. The polarity_scores function uses the Valence Aware Dictionary and Sentiment Reasoner (VADER) [1], which covers lexical items labeled with polarity valence related to the semantic orientation, including a distinction between negative items (”couldn’t”, ”daren’t”, ”shouldn’t”, ”nowhere”, ”despite”) and boosting items (”completely”, ”entirely”, ”unbelievably”, ”little”, ”partly”).

5.2 Feature Selection (***Tudor***)

Having selected a wide range of features that are not equal with respect to shape - as they depend on the structure and the size of each tweet - we computed the mean and the standard deviation for the length, and sentiment (positive, negative) of the tweets for each individual user.

Based on the grammatical information acquired through POS tagging, we count the occurrences of personal pronouns, proper nouns and verbs per user. Buntain & Golbeck [6] found personal pronouns to be particularly informative in regards to fake news detection. We hypothesise that proper nouns and verbs are informative of fake news under the assumption that spreaders tend to discuss a specific event or a particular person in combination with *call to action* verbs such as ”watch”, ”click”, ”follow”. However, upon further data exploration, we learned that both authentic and malicious users tend to use a relatively equal amount of proper nouns and verbs.

In regards to the TF-IDF vectoriser, we set and tested different values [100, 300, 500, 1000] for the maximum number of features to be included in the feature space. The best performing TF-IDF feature matrix is based on 500 words. Similarly, we ignore terms that exceeded a maximum threshold (50%, 70%, 90%) for number of documents that the term appeared in. In the final implementation, the threshold is set to 70%. Additionally, we experimented with uni-grams and bi-grams. We also removed high-frequency stop-words (such as ’and’, ’in’, ’the’ etc.) using the NLTK library.

5.3 Models (*Tudor*)

To establish a baseline, we utilise the $DummyClassifier$ module from sklearn. The technique we employed is called $most\_frequent$ and works by predicting the most frequent class in the training set for all values in the test set. This provided us with a better understanding of the nature of the dataset and the proportionality between fake news spreaders and non-fake news spreaders.

Random Forest is a common classification algorithm, frequently used due to its ability to yield stable results over various, multifaceted datasets with only small adjustments. We use the $RandomForestClassfier$ module, provided by sklearn. In order to optimise performance, we experimented with different numbers of decision trees [100, 300, 500, 1000] and discovered that 500 estimators yielded the best overall performance. In addition, we fused the results from multiple random forests, trained on different features.

Gradient Boosting is an ensemble of weak learners for data classification. This algorithm tends to perform better than Random Forests [19]. We utilised the $GradientBoostingClassifier$ module from sklearn. For Gradient Boosting algorithms, the number of estimators plays an important role, as the decision trees are combined in an additive manner with the purpose of reducing the loss of the previous step. Implying a proneness to amplify errors from previous branches. On this basis, we experimented with multiple values and ended up using 500 estimators, similar to the previous model.

Finally, we implement a dense neural network, using TensorFlow and the Sequential model included in the Keras package. After experimenting with different numbers of Dense hidden layers [1, 2, 3] and activation functions, the final implementation posses the following algorithmic architecture: (1) 1 Dense layer with 64 neurons and ReLU activation function, (2) 1 Dropout layer with 30% dropout rate, (3) 1 Dense layer with 32 neurons and ReLU activation function, and (4) 1 output layer with neuron and sigmoid activation function. The model is compiled with the Adam optimiser and the binary cross-entropy loss function. During the training phase, we split the data into batches of 10 users and run the model for 50 epochs.

For evaluation, we opted for a 5-fold Stratified Cross Validation scheme. What motivated this particular choice for validation is the relatively small size of the dataset. In this regard, each fold should include an approximately equal number of fake and authentic news spreaders.

6 Results (*Everyone*)

In this section, we present the results from our experiments in order to determine and show which of our approaches performed best on the task of classifying fake news spreaders based on the provided dataset. First, we present the results from the English tweets, followed by the results obtained from using the Spanish tweets.

In the following tables, by ”POS counts” we refer to the total number of personal pronouns present in all tweets per user. We use ”sentiment” to refer to the mean and standard deviation of the positive and negative sentiments computed across all tweets for a specific user. By structural features, we intend the mean and standard deviation computed across the tweet lengths for a certain user.

6.1 English (*Tudor*)

The first table (Table 1) provides an overview of the different models tested with different combinations for feature extraction. We get a baseline mean accuracy of 50% for the $DummyClassifier$. The best performing model, with 74.66% mean accuracy over 5 folds, is the Gradient Boosting implementation with the TF-IDF vectoriser as feature extraction method. Importantly, the embeddings are uni-grams, as they yielded the best results.

Model	Features	Mean Accuracy
Dummy Classifier	Most Frequent	50.00%
Random Forest	POS counts, sentiment, structural features	63.66%
Gradient Boosting	POS counts, sentiment, structural features	62.00%
Random Forest	TF-IDF	68.66%
Gradient Boosting	TF-IDF	**74.66%**
Neural Network	TF-IDF	70.00%
Neural Network	Doc2Vec	67.66%

Table 1: Performance of classifiers using various text-feature combinations

The table below (Table 2) includes results gathered from implementing early and late feature fusion. As expected, early fusion does not improve the previously outlined results, as both the dimensionality and the magnitude of the manually extracted features and the TF-IDF features differ to a great extent. On the other hand, late feature fusion performs better on English text, which denotes that adding semantic information positively influences the outcome of the algorithm.

Model	Features	Fusion	Mean Accuracy
Dummy Classifier	Most Frequent	-	50.00%
Random Forest	TF-IDF, sentiment	Early	71.33%
Gradient Boosting	TF-IDF, sentiment	Early	74.66%
2 x Random Forest	TF-IDF, POS counts, sentiment, structural features	Late	66.66%
2 x Grandient Boosting	TF-IDF, POS counts, sentiment, structural features	Late	70.00%
GB and RF	TF-IDF, POS counts, sentiment, structural features	Late	**75.33%**

Table 2: Performance of classifiers using late and early feature fusion

6.2 Spanish (*Jonas*)

Below we present the results from applying our methods to the Spanish data. We see that overall models performance improved for the Spanish text data with a mean accuracy score of 77.6% (Table 3)). For Spanish, using both uni-grams and bi-grams improved performance. The results demonstrate the generalisability of our framework onto a Latin language.

Model	Features	Mean Accuracy
Dummy Classifier	Most Frequent	50.00%
Random Forest	POS counts, sentiment, structural features	69.66%
Gradient Boosting	POS counts, sentiment, structural features	66.33%
Random Forest	TF-IDF	75.66%
Gradient Boosting	TF-IDF	73.66%
Neural Network	TF-IDF	**77.00%**
Neural Network	Doc2Vec	69.00%

Table 3: Performance of classifiers using various text-feature combinations

We were surprised to discover that appending sentiment information to the TF-IDF matrix early on in the prediction process has a positive impact on our model's ability to discriminate fake news spreaders (Table 4).

Model	Features	Fusion	Mean Accuracy
Dummy Classifier	Most Frequent	-	50.00%
Random Forest	TF-IDF, sentiment	Early	**77.66**%
Gradient Boosting	TF-IDF, sentiment	Early	74.33%
2 x Random Forest	TF-IDF, POS counts, sentiment, structural features	Late	76.66%
2 x Grandient Boosting	TF-IDF, POS counts, sentiment, structural features	Late	74.00%
GB and RF	TF-IDF, POS counts, sentiment, structural features	Late	76.00%

Table 4: Performance of classifiers using late and early feature fusion

7 Discussion (*Everyone*)

In this section, we outline reflections and considerations regarding our choice of method. Further, we bring attention to several possible limitations and drawbacks, as well as suggestions for how these could have been addressed.

Next, we evaluate the dataset, and what implications it has had for the results. Then, we proceed to compare how the English and Spanish tweets performed, their different results and possible ideas to why they performed the way they did. In addition, we contemplate on the complications misclassification bias might pose to our model, expanding further by considering other points of critique. Lastly, we reflect on future work and what we could have done as logical next steps in the process, given that we had more time and resources available.

7.1 Reflections on Methods (*Helena/Tudor*)

As an initial step in the process of inquiry, we set out to identify a user, or an account that is associated with fake news spreading. With an accuracy score exceeding 74% for all tests, our classifiers demonstrate classification performance that outperforms that of our baseline model considerably. This supports the hypothesis that Machine Learning can be used to identify fake news spreaders in the Social Media context.

In our endeavour to understand what characterises a malicious account, we learned that multiple personal pronouns combined with sentiment values are features that are indicative of fake news spreaders [23]. These findings are similar to the results presented in [18].

For the final implementation, we decided to make personal pronouns the sole lexical POS feature. We speculate personal pronouns might be frequent in tweets that aim to misinform by drawing on polarity, such as communicating a sense of community or contention between fractions; (”we” and ”us”, versus ”we” and ”them”).

In addition to considering personal pronouns, we also experimented with proper nouns. However, we learned that these decreased classification accuracy. Some examples of tweets from spreaders labeled as fake include proper nouns such as: ”Trump”, ”Meryl Streep”, ”Jordan Peterson”, ”Texas” and ”Jean Claude Van Damme”. However, non-fake news tweets also include proper nouns such as: ”Walmart”, ”Tom Ford”, ”Jacinda Ardern”, and ”New Zealand”. Thus, we believe that the proper noun tag in itself might not contribute substantially to the task of deciphering between fake news and authentic news.

We tested removing upper casing and hashtags but these additions resulted in decreased performance. Quite reasonably, this may be due to properties particular to the specific data used for this work. For other social media data, excluding hashtags might render itself to be a good strategy.

A possible drawback with our framework is that it does not account for semantic meaning. Previous work [25] shows that semantic properties are informative with respect to misinformation recognition. We are dealing with text data, which holds semantic information. In its current form, our feature extraction methods do not ”read” the tweets as such however - they analyse them purely based on lexical and sentiment features. Our framework also fails to capture synonyms, antonyms, polysemous words, as well as sets of semantically interrelated words such as government/ president/ state. The inclusion of more fine-grained semantic features might have resulted in a more rigorous and more accurate user recognition classifier.

Similar comments could be made about sentiment. Aside from measuring sentiment polarity based on the VADER dictionary, previous work has demonstrated that contribution from other dictionaries, such as the psycho-linguistic Linguistic Word Count (LIWC), dictionaries for moral terms, or biased expressions improves the classification of suspicious news [25].

7.1.1 Limitation to the Dataset (*Helena/Tudor*)

We believe the dataset does not include sufficient information about the individual users in order for us to properly built highly accurate machine learning classifiers. The credibility of a user on social media platforms shows strong correlation to the number of connections in the network, the account status (verified / not verified), the post frequency and the amount of reactions (likes, shares) a post triggers. This is not information we have at our disposal.

7.1.2 English Tweets vs Spanish Tweets (*Tudor*)

As mentioned previously, we made adjustments to the feature extraction methods in order to accommodate possible language dissimilarities. The models based on TF-IDF perform better with uni-gram encoding for English tweets. Whereas the Spanish fake news spreaders - in contrast - are better identified by incorporating both uni-grams and bi-grams.

We suspect this incongruence to be related to the lexical differences between the English and Spanish language in general. In order to get a more holistic understanding of the problem, we computed the 10 most informative TF-IDF features for both English and Spanish tweets using Pearson's χ^2 test. In English tweets, the name of the president of the USA (Figure 1) is highly likely to be associated with fake news tweets. On a similar note, call to action verbs such as ”watch”, are ranked as good descriptors of misinformation. For the Spanish data, the word ”hashtag” is recurrent among the top 10 TF-IDF best informative features for misinformation, as seen in the figure below (Figure 2)

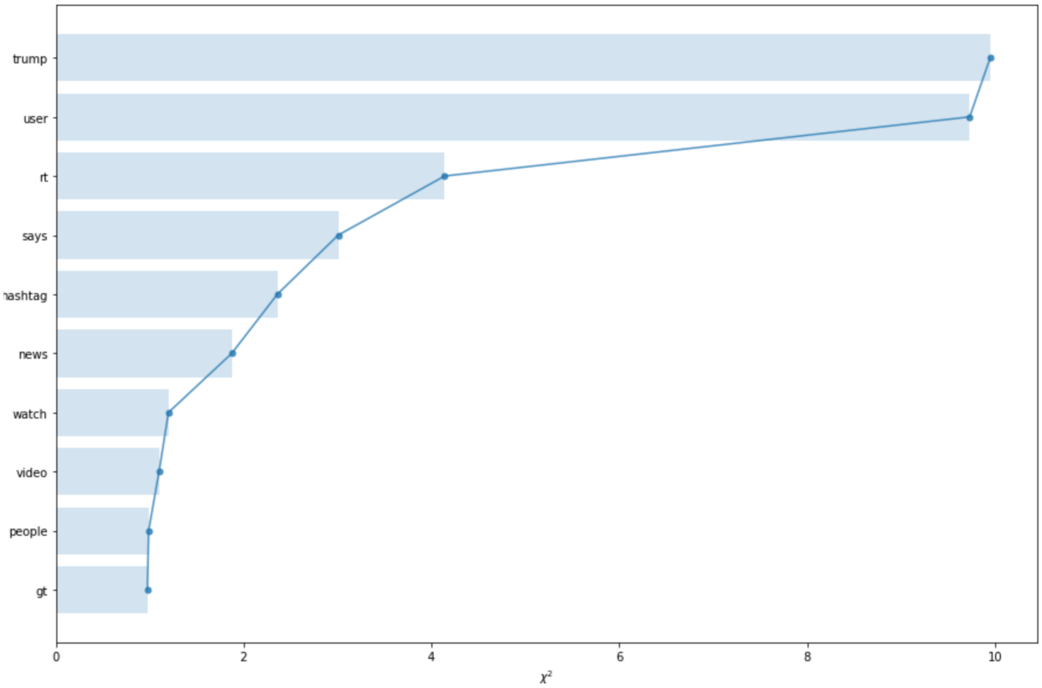

Figure 1: Figure shows the χ^2 scores for the 10 highest rated words in ascending order for English tweets.

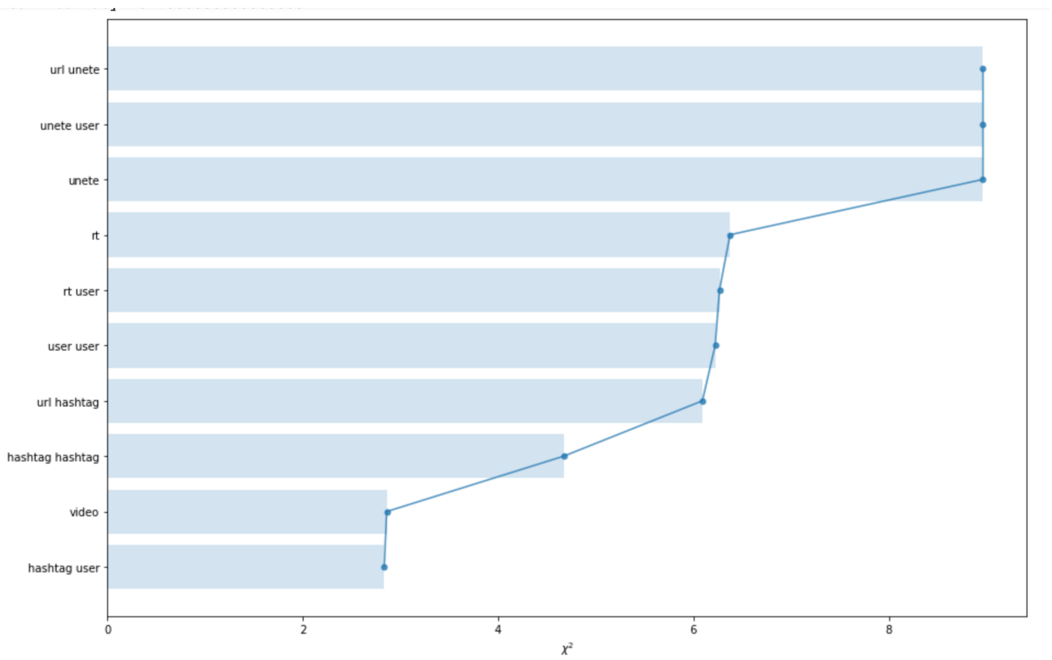

Figure 2: Figure shows the χ^2 scores for the 10 highest rated words in ascending order for Spanish tweets.

7.2 When Machines Fails (*Jonas*)

When evaluating and considering our results, we achieved accuracies in the 74-77 % range for Spanish and English tweets. While these accuracies are impressive in of themselves, they do entail some considerable margin of error. The potential outcome of this error is that real news will get tagged as fake news, and that fake news will slip through the cracks and be deemed as real news. In essence, we are dealing with the issue of false positives and positive negatives in relation to a possible misclassification bias [8, 26].

With implementation methods like variations to classification approaches, people put trust in their outcomes. However, within their margin of error, they carry the ability to scrutinise and undermine real, trustworthy resources. In similar fashion, the misclassification bias problem can potentially pose an equally disturbing and pernicious threat as the phenomenon of fake news did to begin with. In this way, the counter-measure of implementing a fake news detector has generated a new and equally baleful problem in its endeavour to rid the world of propaganda and fake news.

It should be noted however, that while this is a real cause of concern when having an accuracy of roughly 75%, the problem in general will be of lesser concern as new classification methods develop and become more precise.

7.3 Critical Reflections on Scope (*Helena*)

The dataset that we have been working with for the purposes of this project is fake news specific. Given the similarities between fake news, rumour and propaganda that were discussed previously, there is a risk that our framework is not sufficiently fine-grained to distinguish between these differences on novel data. That said, it is nevertheless the case that despite the concepts being distinct, the surface properties and patterns may still be similar so that identifying linguistic features informative of misinformation might suffice to discriminate fake news from authentic news if the dataset is target/news specific - i.e. extracted tweets reporting seemingly factual events in contrast to tweets presenting a position in favour of a political doctrine. Previous works have constructed datasets with tweets relating to a specific event, such as a crisis and then proceeded to extract data specific features from that [3] [12].

This challenge should be further viewed in light of a challenge previously raised in the discussion. If our features are coarse, and our classifiers consequently broad, there is a risk that propaganda might be classified as fake news, or that authentic news would be predicted as inauthentic. This point further relates to the broader issue of generalisability, stemming from the inherent limitations of the representational properties of labeled data. The labeled data presents an idealisation, a representation of a subset of fake versus authentic news posts tweeted in a given time period, and revolving around a set of topics. A classifier trained and tested on this data may yield higher accuracy than if tested on a different dataset or on data *in the wild.* Discourse changes with time and context, and different social networks may display particular distinctive patterns. We also speculate that the topic of interest in a set of fake news tweets may also impact the choice of stylometric and other features, fake news tweets about purely political affairs possibly being different from fake news tweets about about global warming [9].

7.4 Future Work (*Everyone*)

Our framework does not incorporate much semantic information. Future work would benefit from including semantic features. One strategy would be to use word or sentence encoders that use dictionaries for transfer learning such as ELMo or BERT. This problem is deeper than merely establishing that non fake news spreaders use shorter words, or number of words etc. Future work should address what is actually being said.

One suggestion for future endeavours is to implement Word2Vec embeddings using a model tailored for social media content. The features could be further processed with a Recurrent Neural Network model, such as LSTM with long term memory, in order to gain contextual information when analysing semantic information [15].

A potentially fruitful strategy could be to accompany TF-IDF with additional information from various annotated corpora. We've seen that emotional, moral and bias cues can improve prediction performance on inauthentic news data [25].

We suggest future work combine lexical, statistical and sentiment features with features for network structure and network behavior cues by counting retweets or mapping network structures [25] [24].

8 Conclusion (*Everyone*)

With this project, we set out to test whether we can accurately distinguish between fake news spreaders and spreaders of authentic information using machine learning methods in an online social media context. While we do acknowledge that our solution could be improved upon, our achieved results for both English and Spanish Twitter data support our hypothesis, as classification on data from both languages outperforms the baseline model.

9 Questions and Answers (*Everyone*)

1. Depending on the topic in question, we would attempt to include a set of known facts about it and end the post on a fake note. In accord with the Recency Effect, we speculate that people best remember the last bits of information they are presented with [13, 21]. The mix of facts and fake news would hopefully confuse the algorithm and get our post into a gray area of uncertainty. Another key aspect is using human-like text, with proper syntax and grammar. Even though writing with capital letters is tempting, as it might engage a larger audience. We would keep everything lower case - proper nouns and beginnings of sentences aside. Avoid exaggerated sentiment polarity - balance statements so as not to be overly positive or negative. Avoid heavily value laden semantics with words such as ”loyal”, ”betrayal”, ”bad”, ”traitor”. Here is an example of a fake tweet we would not post: ”X betrayed Y country by start the CORONAVIRUS pandemic!!!”.

   Essentially, what we gathered from our analysis was; that (1) multiple pronouns are a traits of fake news. So we best to avoid that. (2) Sentiment valence is similarly a good indicator of a fake news post. Therefore, write in a very neutral tone. As far as it is possible, try to avoid certain keywords (3) like ”Trump” etc., as these are more likely to be in fake news posts and would therefore increase the likelihood of being detected. We should avoid (4) call to action verbs, as these too are key components in many fake news posts.

   In this way, by considering how our implementation functions, it is interesting to discover and understand how to deceive the algorithm into believing a fake post is actually legitimate by learning what will make it believe a post is fake news.

   In social media, the reach of a given message is highly dependant on the properties of the network, or the person spreading it. Reasoning from this fact, we would then first create a set of connections with people of different backgrounds. This would be done to ensure that our information spreads diversely and exponentially as we keep on persuading real users. Homogeneous networks seem to be predictive of fake news spreading - circulating similar information among the same or interconnected groups of users.

2. Given unlimited data and processing power, we would direct our focus towards building a model, that is able to capture contextual information with the hope of identifying topics trending in fake news at any moment in time. Essentially, this would create more value for social media platforms and the public at large, as it would stop fake news contributors from interfering with elections, or other facts and events globally discussed.

   In order to capture contextual information and essentially understand the ideas behind a tweet, we would make use of Bidirectional Long Short Term Memory Networks (LSTM) - mapping long term dependencies as opposed to for instance conventional RNNs. Under normal conditions, a drawback to LSTM networks is that they are based on a high number of trainable parameters (4x more than ordinary RNNs). Given our imaginary setup, we believe that it would be able to capture unforeseen descriptors for classifying fake news contributors.

   Another relevant contextual aspect would be looking at user and social network behaviour. This could be approached by using a PageRank algorithm - with the intent of mapping and studying structural retweeting patterns, temporal patterns of information spreading, or similar.

   Additionally, we would do source checking, and look at comment sections. As we speculate that inauthentic news tweets might be more commonly questioned than authentic news.

   Yet another possibility would be to use a two-step procedure by first, identifying ”suspicious” data that displays misinformation markers. Then, extract fake news specific posts/contributors for more fine-grained and on-target classification.

   For feature extraction, we would expand on semantic properties - using e.g. BERT or some similar state-of-the-art word or sentence encoders. In addition, methods such as TF-IDF can be accompanied by utilising corpora annotated for emotion and persuasion expressions, bias expressions, as well as moral and value laden expressions. As previous research has shown, such cues to be informative about suspicious news (see e.g. [25]). In addition, topic modeling may contribute to zooming in on fake news as there may be some bias with respect to topics covered inauthentic news posts.

10 Bibliography

References

[1] nltk.sentiment.vader — NLTK 3.5 documentation.

[2] Facebook and Google pledged to stop fake news. So why did they promote Las Vegas-shooting hoaxes?, Oct. 2017. Library Catalog: www.latimes.com Section: Business.

[3] S. Antoniadis, I. Litou, and V. Kalogeraki. A Model for Identifying Misinformation in Online Social Networks. pages 473–482, Oct. 2015.

[4] C.-S. Atodiresei, A. Tănăselea, and A. Iftene. Identifying Fake News and Fake Users on Twitter. *Procedia Computer Science*, 126:451–461, Jan. 2018.

[5] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, Mar. 2018.

[6] C. Buntain and J. Golbeck. Automatically Identifying Fake News in Popular Twitter Threads. *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215, Nov. 2017. arXiv: 1705.01613.

[7] G. L. Ciampaglia. Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science*, 1(1):147–153, Jan. 2018.

[8] J. Davis and S. O'Flaherty. Assessing the accuracy of automated twitter sentiment coding. 16:35–50, Jan. 2013.

[9] M. Del Vicario, W. Quattrociocchi, A. Scala, and F. Zollo. Polarization and Fake News: Early Warning of Potential Misinformation Targets. *arXiv:1802.01400 [cs]*, Feb. 2018. arXiv: 1802.01400.

[10] B. Ghanem, P. Rosso, and F. Rangel Pardo. Stance Detection in Fake News: A Combined Feature Representation. Nov. 2018.

[11] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan. False information detection in online content and its role in decision making: a systematic literature review. *Social Network Analysis and Mining*, 9(1):50, Sept. 2019.

[12] S. Hamidian and M. Diab. Rumor Identification and Belief Investigation on Twitter. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 3–8, San Diego, California, June 2016. Association for Computational Linguistics.

[13] J. Hartono. The Recency Effect of Accounting Information. *Gadjah Mada International Journal of Business*, 6(1):85–116, June 2013. Publisher: Universitas Gadjah Mada.

[14] A. Jewett. Detecting and Analyzing Propaganda. *The English Journal*, 29(2):105–115, 1940. Publisher: National Council of Teachers of English.

[15] K. Jiang, S. Feng, Q. Song, R. A. Calix, M. Gupta, and G. R. Bernard. Identifying tweets of personal health experience through word embedding and LSTM neural network. *BMC Bioinformatics*, 19(8):210, June 2018.

[16] S. Krishnan and M. Chen. Identifying Tweets with Fake News. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 460–464, July 2018.

[17] P. Kumar and S. Schoenebeck. The Modern Day Baby Book: Enacting Good Mothering and Stewarding Privacy on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1302–1312, New York, NY, USA, 2015. ACM.

[18] J. Li, Z. Ye, and L. Xiao. Detection of Propaganda Using Logistic Regression. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 119–124, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[19] J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, 5(3):S11, May 2011.

[20] M. Orlov and M. Litvak. Using Behavior and Text Analysis to Detect Propagandists and Misinformers on Twitter. Sept. 2018.

[21] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 50–57, Istanbul, Turkey, Nov. 2014. Association for Computing Machinery.

[22] T. Reddy, V. v. Bulusu, and V. Reddy. A survey on Authorship Profiling techniques. 11:3092–3102, Mar. 2016.

[23] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, Jan. 2003.

[24] M. D. Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, Jan. 2016. Publisher: National Academy of Sciences Section: Physical Sciences.

[25] S. Volkova, K. Shaffer, J. Jang, and N. Hodas. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. pages 647–653, Jan. 2017.

[26] C. T. Weber and S. Syed. Interdisciplinary optimism? Sentiment analysis of Twitter data. *Royal Society Open Science*, 6(7):190473, July 2019.

[27] A. Weston. *A Rulebook for Arguments.* Hackett Publishing Company, Incorporated, Cambridge, 2018. OCLC: 1027200473.

[28] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, and S. Havlin. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9(1):1–14, Dec. 2020. Number: 1 Publisher: SpringerOpen.

[29] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Towards Detecting Rumours in Social Media. Apr. 2015.

[30] O. Çıtlak, M. Dörterler, and A. Doğru. A survey on detecting spam accounts on Twitter network. *Social Network Analysis and Mining*, 9(1):35, July 2019.