From Fake News to Great Rules

Using Machine Learning to create human-interpretable rules to spot fake news

Jonas Malm

TDDE16
January 2021

1 Abstract

Fake news is a big problem in society: it influences elections and creates vaccine hesitation, while most people struggle to identify fake news better than chance. This paper introduces a novel approach to help people make this distinction, by clustering the most influential words in a Naïve Bayes classifier based on similarity to create human-interpretable rules to identify fake news. Furthermore, through testing combinations of the clusters three simple rules are proposed which could significantly help humans differentiate fake news from real news with an estimated F1-score of >75%.

2 INTRODUCTION

Fake news is defined as fabricated information imitating traditional news content, although the editorial norms for making sure the information is credible is not present in contrast to a traditional news organization (Lazer *et al.*, 2018). Fake news is spread more widely in social media than in mainstream news (Balmas, 2014) and a 2016 study showed that over 60% of adults source their news from social media and that the share is increasing (Gottfried and Shearer, 2019), making distinguishing between fake news and real news more important than ever.

Fake news wields a significant political power: in a recent study of the 2018 Italian election, exposure to fake news was shown to favor populistic parties (Cantarella, Fraccaroli and Volpe, 2020). Furthermore, several studies have shown that the primary cause of vaccine hesitancy is fake news and misinformation on social media (Jolley and Douglas, 2014; Dubé, Vivion and MacDonald, 2015; Aquino *et al.*, 2017), making it an imperative for governments to equip its citizens with tools to identify fake news.

Generally, most people find it challenging to identify fake news. One study found that only 17% of participants could correctly identify fake news better than chance (Moravec, Minas and Dennis, 2018), while a large survey of 3015 US adults found that 75% of participants assessed fake news as accurate (Silverman and Singer-Vine, 2016).

Thus, the aim of this project will be to apply text mining on a large dataset of fake and real news in order to attempt to extract rules that can be used by humans in order to more accurately identify fake news.

3 THEORY

In this section, algorithms and processes used in the project will be explained.

3.1 NAÏVE BAYES

Since the aim of the project is to create human-interpretable rules for identifying fake news, Naïve Bayes is a good algorithm to use since it bases its predictions on individual words or ngrams. From the trained Naïve Bayes model we can extract the individual likelihoods, e.g. p("mainstream" | Fake).

3.1.1 The Naïve Bayes classifier

Naïve Bayes is a probabilistic classifier, meaning that the posterior probability will be calculated for all possible classes C and the classifier will predict the class c with the highest posterior probability denoted by ĉ (Jurafsky and Martin, 2021), as seen in Equation 1. The classifier is derived from Bayes' rule (Stone, 2013) as seen in Equation 2, and for the classification application the denominator can be removed as this term is constant for all classes c (Jurafsky and Martin, 2021).

Equation 1: Naive Bayes classifier, adapted from (Jurafsky and Martin, 2021)

$$\hat{c} = argmax_{c \in C}\{p(c \mid x)\} \propto argmax_{c \in C}\{p(c)p(x \mid c)\}$$

Equation 2: Bayes' rule, adapted from (Stone, 2013)

$$p(c \mid x) = \frac{p(c)p(x \mid c)}{p(x)}$$

3.1.2 Naïve Bayes for text classification

When applying Naïve Bayes for text classification, there are two underlying assumptions: (i) the order of words does not matter and (ii) the probabilities $p(x_i|c)$, where each feature x_i is a word, are independent for a given class c (Jurafsky and Martin, 2021). The likelihood can then be written as in Equation 3.

Equation 3: The effect of the Naive Bayes assumption on the likelihood function, adapted from (Jurafsky and Martin, 2021)

$$p(x \mid c) = p(x_1, x_2, \ldots, x_n \mid c) = p(x_1 \mid c) * p(x_2 \mid c) * \ldots * p(x_n \mid c)$$

The calculations are done in log space which accomplishes two things: (i) speed is increased since addition is faster than multiplication and (ii) underflow is avoided (Jurafsky and Martin, 2021). By applying the log function to the predictor and factoring in Equation 3, Equation 4 is derived.

Equation 4: The Naive Bayes classifier in log space, adapted from (Jurefsky and Martin, 2021)

$$\hat{c} = argmax_{c \in C}\{logp(c) + \sum_{i} logp(x_i \mid c)\}$$

3.1.3 Training the Naïve Bayes classifier

In order to use the classifier in Equation 4, we need to estimate p(c) and $p(x_i|c)$.

The prior, p(c), is estimated according to Equation 5, where N_c is the number of documents with class c and N is the total number of documents (Jurafsky and Martin, 2021).

Equation 5: Estimation of the prior, adapted from (Jurafsky and Martin, 2021)

$$\hat{p}(c) = \frac{N_c}{N}$$

To estimate $p(x_i|c)$, the probability of observing word x_i in a sentence of class c, Equation 6 is used (Jurafsky and Martin, 2021). Count(x_i, c) is the number of times the word x_i is observed in a document of class c, V is the vocabulary i.e. the set of all words in the documents, and α is a smoothing factor (Jurafsky and Martin, 2021).

Equation 6: Estimation of p(xi|c), adapted from (Jurefsky and Martin, 2021)

$$\hat{p}(x_i \mid c) = \frac{count(x_i, c) + \alpha}{\sum_{x \in V}(count(x, c) + \alpha)}$$

According to Jurafsky & Martin (2021) α is included since without it $p(x_i|c)$ would be zero if x_i does not exist in a document with class c. Since these probabilities are multiplied or its log is added to estimate p(c|x), this would mean the entire probability p(c|x) would be zero if just one word is missing in documents with class c. α is usually set to 1 in text mining applications (Jurafsky and Martin, 2021).

3.2 TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

A problem with using counts to estimate $p(x_i|c)$ is that it assumes that a word has a stronger predicting power just because it occurs more frequently (Jurafsky and Martin, 2021). According to Jurafsky & Martin (2021), raw counts are not discriminative when judging which words are more important in a document. A common solution for this problem in text mining applications is using term frequency-inverse document frequency. This often complements the pre-processing step of removing stop words, which is defined as words that are very frequent and hold little to no informational value, such as 'the' and 'a' (Jurafsky and Martin, 2021).

3.2.1 The tf-idf weighting scheme

Assuming N documents and that term x_i occurs in n_i of these documents, the inverse document frequency or idf of term t_i can be defined according to Equation 7 (Robertson, 2004).

Equation 7: Basic idf, adapted from (Robertson, 2004)

$$idf(x_i) = log\frac{N}{n_i}$$

The term frequency-inverse document frequency, or tf-idf, is then defined according to Equation 8, where tf(x_i, d) is the number of occurrences of term x_i in document d (Jurafsky and Martin, 2021).

Equation 8: Calculating the tf-idf weight, adapted from (Jurefsky and Martin, 2021)

$$tfidf(x_i, d) = tf(x_i, d) * idf(x_i)$$

In scikit-learn, the python package which will be used to calculate the tf-idf weights, the idf is smoothed such that for terms where $n_i = N$ will not be completely ignored as well as preventing divisions with zero illustrated in Equation 9 (scikit-learn developers, 2021b). This smoothed idf weighting scheme will be used in this study because of the aforementioned benefits.

Equation 9: idf weights implemented in scikit-learn, adapted from (scikit-learn developers, 2021)

$$idf(x_i) = log\left(\frac{N+1}{n_i+1} + 1\right)$$

3.2.2 Using tf-idf in Naïve Bayes

Replacing the raw frequency measure used to estimate the probability of observing word x_i in a sentence of class c, $p(x_i|c)$ from Equation 6, with the tf-idf measure Equation 10 is derived.

Equation 10: The Naive Bayes classifier using the tf-idf weighting scheme

$$\hat{p}(x_i \mid c) = \frac{\sum_{d:class(d)=c} tfidf(x_i, d) + \alpha}{\sum_{x_i \in V}\left(\sum_{d:class(d)=c} tfidf(x_i, d) + \alpha\right)}$$

The estimator in Equation 10 in combination with the idf weighting scheme from Equation 9 will be used in this study.

3.3 Word embeddings and vector space models

Word embedding is the act of representing words as vectors, the most simple version being the one-hot encoding where each word in a vocabulary is associated to an index, in a vector the length of the vocabulary, which is set to one while the others indexes are set to zero (Pilehvar and Camacho-Collados, 2020). As an illustrative example, suppose our vocabulary consists of 'hello', 'my', 'friend': then the one-hot encoding for 'hello' could be (1, 0, 0) and 'my' (0, 1, 0).

However, one-hot encoding has a few major drawbacks: similarity is very hard to measure and a large vocabulary requires an equally large sparse vector to represent a single word (Pilehvar and Camacho-Collados, 2020). This problem was solved by representing words as vectors in a high-dimension continuous space (Pilehvar and Camacho-Collados, 2020). There now exists several methods for embedding words, many originating from generative grammar where each word is represented by semantic features that represent a primitive meaning (Jurafsky and Martin, 2021) as illustrated in Table 1.

Table 1: Example of semantic features, adapted from (Jurafsky and Martin, 2021)

Hen	+ female, + chicken, + adult
Rooster	-female, + chicken, + adult
Chick	+chicken, -adult

Other models are based on latent semantic analysis, where decomposition is applied to a term-document matrix and the first 300 dimensions used as the embedding (Jurafsky and Martin, 2021). More modern models have evolved and pre-trained embeddings are available, which allow the similarity in the meanings of words to be measured (Pilehvar and Camacho-Collados, 2020).

3.4 K-MEANS CLUSTERING

K-Means is the most widely used clustering method, in which k clusters are formed so that the samples contained in the clusters are as similar as possible as defined by a selected distance measure, e.g. Euclidean distance (Aggarwal and Reddy, 2013). The algorithm works as outlined in Table 2.

Table 2: The K-Means clustering algorithm, adapted from (Aggarwal and Reddy, 2013)

Step 1	*Select k items randomly as cluster centers*
Step 2	*Form k clusters by assigning each item to the closest cluster center*
Step 3	*Set the cluster centers to the centroid, i.e. arithmetic mean, of the cluster*
Step 4	*If cluster centers moved less than convergence criteria: EXIT* *Else: Go to step 2*

In the algorithm the number of clusters have to be selected and naturally inertia, i.e. the total within-cluster distance from the cluster center, will decrease as more clusters are added. It is claimed that the optimal number of clusters is found when the slope of the inertia / #clusters curve stops decreasing as rapidly as before, which is known as 'the elbow' (Aggarwal and Reddy, 2013), and a such point is visualized in Figure 1. This point can sometimes be ambiguous, indicating that there are no natural ways to cluster or multiple valid ways (Ketchen and Shook, 1996).

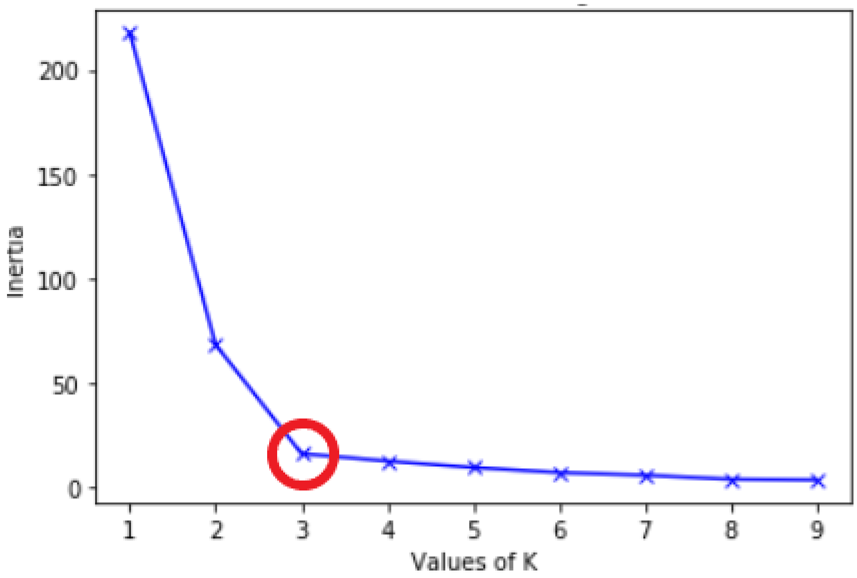
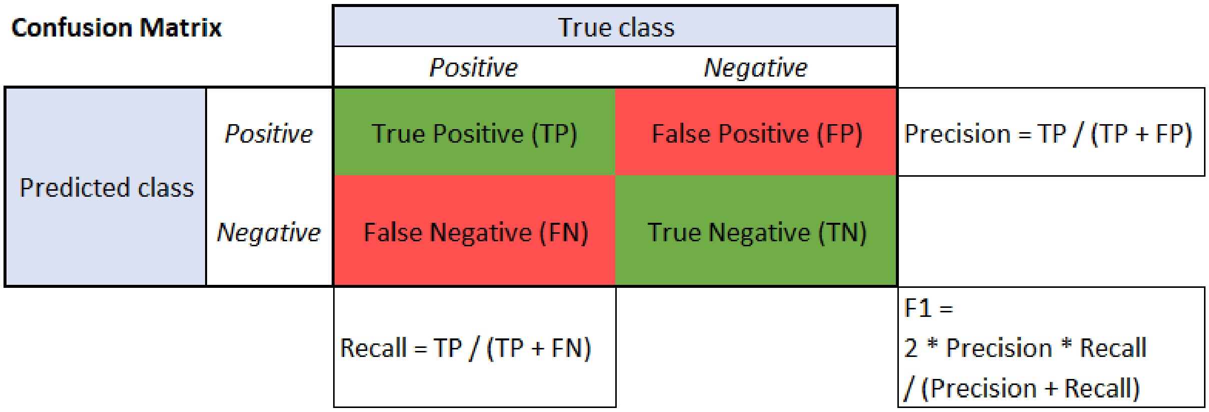


Figure 1: The elbow method using inertia (Geeks for Geeks, 2021)

3.5 Measures of Accuracy

A common basis of measuring accuracy is the confusion matrix, where a classifier's predictions is mapped together with the ground truth (Jurafsky and Martin, 2021) as seen in Figure 2. If the classifier predicts the Positive class for an item but the item's gold label is the Negative class, this is a False Positive and so forth.



Confusion Matrix		True class		
		Positive	*Negative*	
Predicted class	*Positive*	True Positive (TP)	False Positive (FP)	Precision = TP / (TP + FP)
	Negative	False Negative (FN)	True Negative (TN)	
		Recall = TP / (TP + FN)		F1 = 2 * Precision * Recall / (Precision + Recall)

Figure 2: The Confusion Matrix, adapted from (Jurefsky and Martin)

From this definition of TP, FP, FN and TN several measures can be defined. Precision is defined as the share of items predicted as positive that have the positive gold label, while recall is the share of actually positive items correctly identified as positive (Jurafsky and Martin, 2021). There are several F-measures that weight recall and precision together, and the most frequently used is F1 where precision and recall is given equal weight (Jurafsky and Martin, 2021).

4 Data

In order to create rules that will not be only valid for a subset of fake news, time was spent searching for large datasets of fake news. The dataset chosen was the largest dataset not only containing the headline that was found.

4.1 Source

The dataset, WELFake, that will be used is one of the largest publicly available datasets containing fake and real news articles, created by researchers in 2021 by merging the four popular news data sets Kaggle, McIntire, Reuters and BuzzFeed Political (Verma *et al.*, 2021). The authors argue that the mixed sources of the dataset raises the validity of their study. The dataset contains over 72 000 individual articles, of which approximatively 49% are real and 51% fake (Verma *et al.*, 2021).

4.2 Contents

The dataset contains four columns: a unique ID, the title of the article, the text of the article and its gold label where 1 denotes fake news and 0 real news. An example of three of the rows is illustrated in Table 3.

Table 3: Three example rows from the WELFake dataset

ID	Title	Text	Label
21	Hillary's crime family: End of days for the U.S.A	Hillary's crime family: End of days for the U.S.A Based on the foregoing, SOMEONE got to Comey—somewhere along the line By Daily Coin - November 8, 2016 An investor in Dave's fund emailed him asking which way Colorado would vote tomorrow. He replied: "Depends on who counts the votes. I don't believe this is a fair election. I think the Clinton crime machine, with the help of George Soros and a few others, have everything under their control now." […]	1
112	German Police Kill Assailant After Ax Attack Aboard a Train - The New York Times	WEIMAR, Germany — A Afghan youth who came to Germany as a migrant last year attacked several passengers with an ax and a knife on a train in the south of the country late on Monday, injuring at least four people, while 14 others were treated for shock, the police said. After the train made an emergency stop, the attacker fled and was pursued by police officers, who fatally shot him, according to the interior minister of the state of Bavaria, Joachim Herrmann. […]	0
149	"GUATEMALAN" MAN DIES After Falling Into WASTE GRINDER At Meat Plant…Former Worker Claims They Hire "90% Illegal Aliens"…Including "10-12 Yr Old Kids"	This story is absolutely horrific, but not surprising. Bleeding heart liberals aren t doing illegal aliens any favors by fighting for them to come here, only to be abused by their employers, who hire them, and then hold their illegal immigration status over their heads when they demand safer working conditions. (See video below)A worker caught in a machine at a meat processing plant has died in Ohio, according to local authorities.Samuel Martinez, 62, was killed by the machine on Saturday afternoon at the Fresh Mark plant in Canton, Ohio. […]	1

4.3 Pre-processing

Due to difficulties loading the dataset it was preprocessed in Excel before importing the .csv-file into a Pandas DataFrame. The difficulties stemmed from the fact that the file was comma-delimited although both titles and text contained commas. Instead, the delimitation scheme was changed to semicolon-delimitation and all semicolons were replaced with hyphens. This will not affect the study, since these characters will be removed before calculating the tf-idf weights.

Furthermore, the columns Title and Text were joined to form a column named Body. This was done in order to ensure all data present in the dataset would be used for training and prediction.

5 METHODOLOGY

5.1 IMPORTING THE DATA

First, the data is imported into a Pandas DataFrame and the two columns 'title' and 'text' are joined into a new column 'body', which will be used as x-values. The labels are set so that 0 is fake news and 1 is real news, to be used as y-values.

Then the x- and y-values are split into four lists using sklearn's train_test_split: x_train, x_test, y_train and y_test. 70% of the data is used for training and 30% for testing. A large part is reserved for training, since the main goal is to study the trained classifier to extract rules. Lastly, the data is then examined to ensure that the two labels in the datasets are evenly distributed.

5.2 TRAINING A NAÏVE BAYES CLASSIFIER

A simple pipeline is then created consisting of a tf-idf vectorizer and a multinomial Naïve Bayes classifier, both using scikit-learn's classes.

The tf-idf vectorizer is set to ignore English stop words, so that the vocabulary does not contain unnecessary terms with poor predicative power. No bounds are set on minimum document frequency since it would conflict with the goal of finding as many rules as possible, since this would remove some of the words that could be used to find rules. The features are set to only contain 1-gram words, i.e. individual words, since other settings vastly increase the computational power required due to the large dataset. Furthermore, the tokens are not lemmatized as previous studies have found grammatical differences between fake and real news: e.g. verb conjugations (Levi *et al.*, 2019) and frequency of possessive nouns (Horne and Adali, 2017).

The Naïve Bayes classifier is initialized without priors and the smoothing parameter alpha set to one. The classifier is scikit-learn's MultinomialNB, since this classifier is known to perform well on tf-idf vectors (scikit-learn developers, 2021a).

5.3 EVALUATING THE RESULTS FROM THE NAÏVE BAYES CLASSIFIER

Then the trained Naïve Bayes classifier is tested on the testing subset of the dataset and its results are evaluated to ensure the generalizability of the classifier. Since the rules stem from the classifier, the classifier's out-of-sample error should be fairly low in order for the generalizability of the rules to be high.

5.4 CLUSTERING MOST IMPORTANT TERMS AND CREATING RULES

Using the trained classifier, a new DataFrame is created with columns as outlined in Table 4.

Table 4: The columns in the DataFrame used to identify rules

Column name	Definition	Explanation
Word	$A\ word\ w$	A word the classifier has been trained on
Real	$\log p(w\|real)$	The log probability of observing the word in a real news article
Fake	$\log p(w\|fake)$	The log probability of observing the word in a fake news article
Delta	$\log p(w\|real) - \log p(w\|fake)$	The difference between real and fake, i.e. the predicative power of the feature
Importance tfidf	$delta * \sum_{j\ in\ Articles} tfidf(w, d_j)$	Delta multiplied by the sum of the tfidf weights for the word in all articles in the training set; used as a manual reference to importance count
Importance count	$delta * \sum_{j\ in\ Articles} tf(w, d_j)$	Delta multiplied by the total term frequency in all articles in the training set

Since the rules will be based on the logic "if the text contains a word often present in fake articles, it's likely fake", both the frequency of the word and the predicative power i.e. delta of the word is important, the measure Importance Count will be used to measure how important words are. Firstly, a list of the 200 most important words for classifying a text as fake, i.e. the 200 words with the smallest values of Importance Count, is extracted and then the same is done for the 200 most important words for classifying as real.

These words are embedded with the word embeddings in SpaCy's en_core_web_lg library. Then the two embedded lists are clustered separately. The number of clusters is determined with the elbow method.

Each cluster is considered as a separate set of rules, since they now consist of similar words. The words in each cluster is studied to understand what kind of rules they represent. A cluster will be referred to as a rule cluster in the coming sections.

5.5 EVALUATING RULES

To evaluate the rule clusters a classifier is created to implement the rules following the logic in Table 5.

Table 5: The implementation of the rule-based classifier

Step 1	Set the article to lowercase
Step 2	The article receives one negative vote for each word found in the rules and at least once in the article with a fake label
Step 3	The article receives one positive vote for each word found in the rules and at least once in the article with a real label
Step 4	Classify the article as fake if the sum of votes are negative, and real if the sum is positive If there are no votes or the sum of votes is zero the classification is determined by chance

Different combinations of the rule clusters are tested, as outlined in Table 6. Due to restrictions in computational power larger combinations could not be tested. To reduce the computational load, each combination is tested on four different subsets of 1000 articles from the testing data since there are many possible combinations and the testing set is very large. The same four subsets are used to evaluate every combination of rules and the F1-score is calculated for each of the four subsets, which is then averaged to create the final measure as accuracy.

Table 6: The rule cluster combinations tested

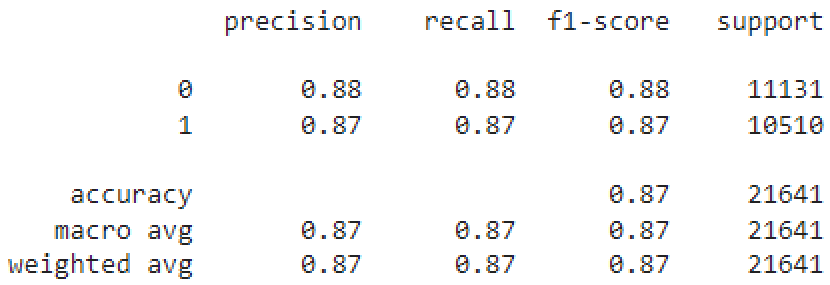
i.	All clusters individually
ii.	One fake and one real cluster
iii.	Two fake and two real clusters
iv.	One fake and two real clusters
v.	Two fake and one real cluster

6 RESULTS

In this section, the result of the project is presented.

6.1 THE NAÏVE BAYES CLASSIFIER

As seen in Figure 3, the classifier performs well on the test data with an F1-score of 87%. The classifier is balanced with very similar values of precision and recall.



```
              precision    recall  f1-score   support

           0       0.88      0.88      0.88     11131
           1       0.87      0.87      0.87     10510

    accuracy                           0.87     21641
   macro avg       0.87      0.87      0.87     21641
weighted avg       0.87      0.87      0.87     21641
```

Figure 3: Results of the Naive Bayes classifier on the testing set

6.2 THE RULE CLUSTERS

In this subsection, the results regarding selection of number, description and combination of clusters to predict are described.

6.2.1 Number of clusters

As seen in Figure 4 and Figure 5, the number of clusters for the real and fake words are set to seven and five respectively using the elbow method. The chosen elbow point is marked by the red circle. Since the plots are somewhat ambiguous, the choice was made with respect to not having too many clusters in order to ensure the generalizability of the rules.

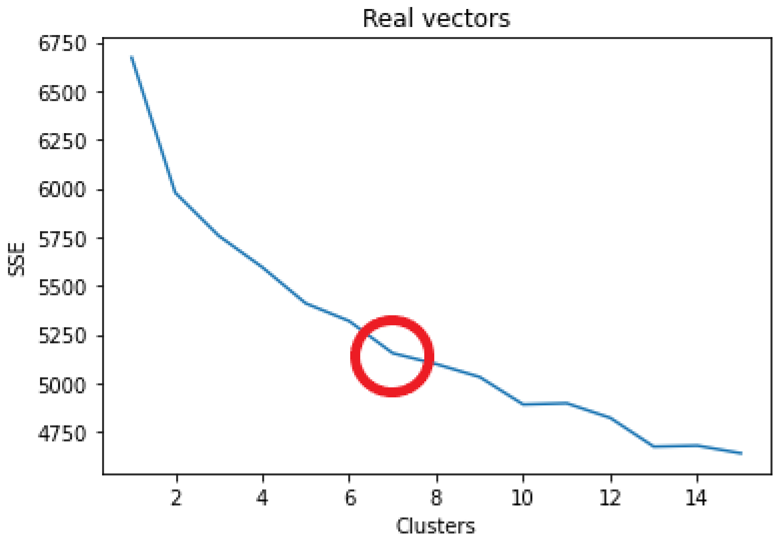


Figure 4: SSE (Interia) by number of clusters for the 200 real words

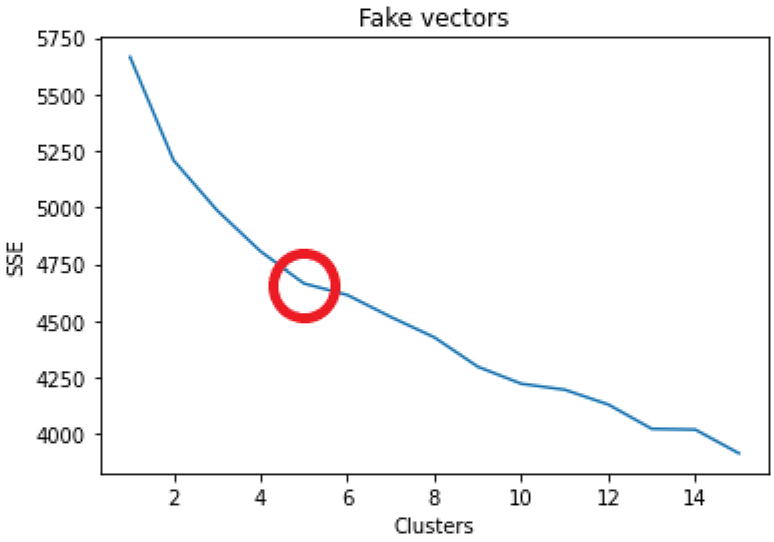


Figure 5: SSE (Interia) by number of clusters for the 200 fake words

6.2.2 Interpreting the clusters

In Table 7 and

Table 8 the clusters are interpreted to summarize their contents. The within-cluster distance is the total distance within the cluster in cluster-space.

Table 7: Fake rule clusters

Cluster	Name	No. words	Within-cluster distance	Example words	Description
0	Common US conspiracy proper nouns	22	122.30	Trump, Hillary, Clinton, Obama, fbi, soros, cia, wikileaks, isis, fox, cnn	People and organizations often present in fake news articles about the US
1	Informal words	111	460.64	Just, like, people, don, didn, watch, apparently, pretty	Abbreviations and words that have a more informal tone
2	Unreliable sources	39	210.41	Video, twitter, facebook, youtube, tweet, images, photo, www, com, https	Words related to information sources that are not credible or plain-text URLs
3	US right-wing fake news words	25	136.69	Mainstream, liberal, flag, propaganda, muslim, conspiracy, gun, nation, racist, cops	Words that are often present in fake news written by the US right-wing
4	Spanish words	3	10.91	Que, la, en	Three Spanish words

Table 8: Real rule clusters

Cluster	Name	No. words	Within-cluster distance	Example words	Description
0	Highbrow political stakeholders	18	87.23	Lawmaker, parliament, senate, opposition, legislation, reform	Non-colloquial words related to politics, mostly stakeholders
1	Highbrow policy words	58	301.99	Trade, agreement, sanctions, budget, government, tax, deal	Non-colloquial words related to politics, mostly policy-related
2	Time and numbers	22	101.98	New, percent, year, week, billion, earlier, expected	Mostly words describing time or numbers, but also a few verbs as ‘called’, ‘saying’ and ‘killed’
3	Geographic words	11	51.29	North, south, capital, border, region, central, city	Words describing a location or geography
4	Titles, weekdays and names	24	122.46	Mr, mrs, ms, Monday, Friday, eu, turkey, Brexit	All seven weekdays, titles, countries and organizations
5	Countries and cities	27	138.76	Washington, China, Korea, Britain, Beijing, European, Moscow	Countries, cities and ethnicities
6	Operational politics	40	198.73	Said, told, officials, statement, committee, department, meeting, union	Words related to day-to-day politics

6.2.3 Best combination of rules

A total of 432 different combinations of rule clusters were tested. Table 9 holds the F1-scores for the 25 combinations which yielded the best F1-score. The full results are presented in Table 14 in the Appendix.

Table 9: The best combinations of rule clusters

Rank	Real clusters	Fake clusters	F1-Score
1	5, 6	0, 2	80,5%
2	5, 6	2, 4	79,9%
3	5, 6	2, 3	79,9%
4	4, 5	2	78,9%
5	4, 6	0, 2	78,8%
6	3, 6	0, 2	78,7%
7	4, 6	2, 3	78,7%
8	2, 4	2, 3	78,5%
9	0, 6	0, 2	78,5%
10	2, 4	2, 4	77,9%
11	3, 6	2, 4	77,8%
12	3, 6	2, 3	77,7%
13	2, 5	2	77,7%
14	1, 5	0, 2	77,6%
15	3, 4	2	77,5%
16	1, 5	2, 3	77,5%
17	0, 6	2, 3	77,4%
18	2, 6	2, 3	77,3%
19	2, 4	0, 2	77,2%
20	4, 5	0, 4	76,8%
21	4, 6	2, 4	76,8%
22	0, 6	2, 4	76,7%
23	6	0, 2	76,6%
24	2, 6	0, 2	76,5%
25	1, 3	2, 3	76,4%

The information in Table 9 is further analyzed in Table 10, Table 11 and Table 12, where frequencies of clusters and cluster combinations in Table 9 are presented.

Table 10: Occurrences of rule clusters in Top 25 combinations

Real cluster	Occurrences in top 25
0	3
1	3
2	6
3	5
4	9
5	8
6	15

Fake cluster	Occurrences in top 25
0	9
1	0
2	24
3	8
4	6

Table 11: Most common same-category cluster pairs in Top 25 combinations

Real cluster pair	Occurrences in top 25
5, 6	3
4, 6	3
3, 6	3
2, 4	3
0, 6	3
4, 5	2
1, 5	2
2, 6	2
2, 5	1
3, 4	1
6	1
1, 3	1

Fake cluster pair	Occurrences in top 25
0, 2	8
2, 3	8
2, 4	5
2	3
0, 4	1

Table 12: Most common different-category rule cluster pairs in Top 25 combinations

Real cluster	Fake Cluster	Occurrences in top 25
6	2	15
4	2	8
5	2	7
2	2	6
6	0	6
6	3	5
3	2	5
6	4	4

7 Discussion

This section will firstly analyze the results to yield rules to identify fake news and then the results will be evaluated and connected to related work.

7.1 Rules for differentiating real and fake news

In this subsection the results will be analyzed in order to extract rules humans can use to identify fake news.

7.1.1 Rules for identifying fake news

The fake rule cluster 2, *unreliable sources*, is present in all but one top 25 combinations which indicates it is one of the most important rules. The 4th, 13th and 15th best combinations feature *unreliable sources* as the only fake rule cluster used. Fake rule clusters 0 (*common US conspiracy proper nouns)*, and 3 (*US right-wing fake news words*) are also prevalent, but only in combination with *unreliable sources* implying their discriminative predicative power is much smaller.

It is quite surprising that the predicative power of fake rule cluster 0 (*informal words*) is so small, as other studies have shown this to be an important feature (Horne and Adali, 2017). The fact that cluster 4 (*Spanish words*) is used in four top 25 combinations is also surprising. This is likely due to peculiarities with the dataset, as it is hard to imagine that Spanish words should entail a higher probability for fake news in general.

7.1.2 Rules for identifying real news

Cluster 6 (*operational politics*) is the most prevalent in the top 25 combinations with, which makes sense as these words are quite advanced and often used by serious news agencies when covering politics. The cluster also contains words related to citing a source, i.e. ‘told’ and ‘said’. The fact that this cluster is effective at predicting could be because large portions of the real articles in the dataset are Reuters articles written on politics. However, it is plausible that using less colloquial and more advanced words infer a smaller probability of the article being fake as noted in previous work (Horne and Adali, 2017).

Clusters 4 (*titles, weekdays and names)* and 5 (*countries and cities*) are also common, but mostly occur in combination with *operational politics* and *unreliable sources*. The predicative power of *countries and* cities could be due to the fact that many Reuters articles start with the location of the reporter, e.g. Weimar, Germany as illustrated in Table 3. In conjunction with *operational politics,* it is reasonable to infer that more advanced and formal highbrow words indicate credibility of the source.

7.1.3 The proposed rules for identifying fake news

Combinations of *operational politics* and *unreliable sources* are the combinations with the best F1-scores: the pair were present in 15 of the 25 best combinations of rules as well as the individual pair being the 27th best combination with an F1-score of 76.4%. Rule cluster *titles, weekdays and names* is also prevalent and has a good predicative power. The rule cluster *countries and cities* could be an effect of a skewness in the dataset and will thus not be included in the recommended rules. Based on this the resulting rules of the project are shown in Table 13.

Table 13: Rules for identifying fake news

Rule	Interpretation
a) Unreliable sources indicate fake news	If the article contains words such as Facebook, Youtube, Twitter, tweet or plain-text URLs the article is likely fake
b) Non-colloquial field-specific words indicate real news	This study showed that non-colloquial advanced political words in political articles indicated that the article is real, and this likely applies to other fields as well
c) Formal referencing indicates real news	Using titles such as Mr, Mrs and Ms; and referencing a specific weekday indicates that an article is credible

7.2 RELATED WORK

One paper studying the linguistics of fake news found that the strongest indicative factor of fake news was the density of agentless passive voice (Levi *et al.*, 2019), i.e. writing ‘the car is being fixed’ instead of ‘the mechanic is fixing the car’. This is concurrent with the results from this paper, as the rule clusters strongly indicating real news contain words which act as agents or imply the use of agents, e.g. the titles ‘Mr’ and ‘Mrs’ in *titles, weekdays and names* as well as ‘officials’, ‘said’ and ‘told’ in *operational politics*.

One paper studying the difference between satire and fake news found that the frequency of first-person singular pronouns, i.e. I, we and us, was the strongest indicative factor of fake news (Levi *et al.*, 2019), which this paper supports as the rule cluster *titles, weekdays and names* consisting mainly of titles was shown to strongly indicate real news and that rule c in Table 13 is valid.

Horne and Adali (2017) studied content-based features of fake, satire and real news articles in three datasets and found that fake news uses a simpler language with less complex words. This also supports the conclusions in this project, as the rule cluster *operational politics* containing advanced words indicated real news. Furthermore, other real indicating rule clusters such *as highbrow political stakeholders* and *highbrow policy words* containing advanced vocabulary alongside the fake indicating rule cluster *informal words* were also formed further supporting the results of Horne and Adali (2017) and that rule b in Table 13 is valid.

Other studies have conflicting results: a study of fake news on Twitter showed that tweets lacking a URL were related to fake news (Castillo, Mendoza and Poblete, 2011) which in conflict with the results of this study as URLs were showed to be a good predictor of fake news. It is however possible that this is specific to the domain of Twitter and not applicable for articles.

7.3 VALIDITY AND LIMITATIONS OF THE STUDY

The authors who created the WELFake dataset also trained a Naïve Bayes classifier, with the same hyperparameters and train-test-split ratio, which yielded an F1-score of 92% (Verma *et al.*, 2021). However, their although their pre-processing was much more advanced with a novel approach consisting of e.g. linguistic feature extraction, selection and set creation; and word embedding. As such it is natural that their classifier performed better, but the F1-score of 87% for the trained Naïve Bayes model in this project indicates an efficient and accurate predictor.

The choice of the number of clusters was somewhat ambiguous, as there were several possible elbow points. Since the resulting clusters were quite uniform, the choices are deemed to be reasonable. However, it is possible that a larger number of clusters could have created rules better adapted to this specific dataset.

Studies have shown that cognitive bias is a large factor in correctly identifying fake news: people are less inclined to think text that align with their beliefs are fake (Moravec, Minas and Dennis, 2018; Cantarella, Fraccaroli and Volpe, 2020). Since this project has not evaluated the rules on real humans, it is possible that the hypothesized positive effects would be nullified by the subjects' cognitive bias and beliefs: i.e. even though a person is aware of the rules, they would hesitate to classify articles aligning with their own beliefs as fake.

With regards to the generalizability of the rules presented, it is possible they are skewed by the large presence of fake news from the US. As evident by the fake word clustering there are two clusters pertaining to US-specific fake news. As such these rules likely do not apply to fake news about other themes or about politics in other countries. However, it is likely that rule b in Table 13 holds for more article categories but this has to be studied further.

Due to the limitations presented, the study should be repeated with a dataset with fake news from outside the US as well as the rules being tested on real people and compared with their political beliefs.

8 CONCLUSION

This paper proposes three simple rules humans can use to identify fake news: unreliable sources indicate fake news, non-colloquial field-specific words indicate real news and formal referencing indicate real news.

Since several studies have shown that average people perform poorer than chance on identifying fake news (Silverman and Singer-Vine, 2016; Moravec, Minas and Dennis, 2018), implying an F1-score of <0.5, it is reasonable to assume that the same people would perform significantly better if they were to make decisions based the rules as outlined in Table 13. Studying the F1-scores of the combinations, it is likely that humans using the rules would perform the task with an F1-score of >75%, although this must be validated by a separate study.

Furthermore, the results presented in this paper supports and builds on previous studies. Levi et al. (2019) found that using agentless passive indicates fake news, while this study found that using agents indicates real news. Horne and Adali (2017) found that using informal words indicated fake news, while this study found that using formal words indicate real news.

9 APPENDIX

9.1 FULL RESULT OF RULE CLUSTER COMBINATION TESTING

Table 14: Full results of rule cluster combination testing

Rank	Real	Fake	F1-score
1	5, 6	0, 2	80,5%
2	5, 6	2, 4	79,9%
3	5, 6	2, 3	79,9%
4	4, 5	2	78,9%
5	4, 6	0, 2	78,8%
6	3, 6	0, 2	78,7%
7	4, 6	2, 3	78,7%
8	2, 4	2, 3	78,5%
9	0, 6	0, 2	78,5%
10	2, 4	2, 4	77,9%
11	3, 6	2, 4	77,8%
12	3, 6	2, 3	77,7%
13	2, 5	2	77,7%
14	1, 5	0, 2	77,6%
15	3, 4	2	77,5%
16	1, 5	2, 3	77,5%
17	0, 6	2, 3	77,4%
18	2, 6	2, 3	77,3%
19	2, 4	0, 2	77,2%
20	4, 5	0, 4	76,8%
21	4, 6	2, 4	76,8%
22	0, 6	2, 4	76,7%
23	6	0, 2	76,6%
24	2, 6	0, 2	76,5%
25	1, 3	2, 3	76,4%
26	0, 4	2	76,4%
27	6	2	76,4%
28	6	2, 4	76,1%
29	4, 5	2, 4	76,1%
30	2, 5	2, 4	76,1%
31	1, 5	2, 4	76,0%
32	5, 6	0, 3	75,8%
33	1	2, 3	75,6%
34	4, 5	3, 4	75,5%
35	0, 1	2, 3	75,4%
36	1, 3	0, 2	75,4%
37	5, 6	2	75,3%
38	0, 1	0, 2	75,2%
39	1, 4	2, 3	75,2%
40	6	2, 3	74,8%
41	2, 5	0, 2	74,8%
42	1	0, 2	74,6%
43	2, 3	2	74,6%
44	6	0, 3	74,3%
45	4, 5	0, 2	74,3%
46	0, 6	0, 3	74,2%
47	2, 5	2, 3	74,2%
48	0, 5	2	74,2%
49	0, 5	0	74,1%
50	4	2	74,1%
51	4, 5	0, 3	74,0%
52	1, 4	0, 2	73,9%
53	3, 6	2	73,8%
54	1, 2	2, 3	73,8%
55	0, 2	2	73,8%
56	2, 6	2, 4	73,8%
57	3, 6	0, 3	73,7%
58	6	0, 4	73,6%
59	5, 6	0, 4	73,6%
60	0, 4	0, 4	73,5%
61	1, 3	2, 4	73,4%
62	5, 6	3, 4	73,4%
63	0, 4	0, 3	73,2%
64	3, 4	3, 4	73,2%
65	3, 5	4	72,9%
66	0, 6	2	72,8%
67	3, 4	0, 4	72,8%
68	2, 5	0, 3	72,7%
69	2, 5	3, 4	72,7%
70	1	2, 4	72,6%
71	0, 1	2, 4	72,6%

72	2	2	72,5%
73	6	3, 4	72,4%
74	1, 2	0, 2	72,3%
75	1, 4	2, 4	72,2%
76	1	2	72,2%
77	3, 4	2, 4	72,0%
78	0, 4	0, 2	71,9%
79	1, 5	0, 3	71,9%
80	4, 5	2, 3	71,8%
81	0, 4	3, 4	71,8%
82	1, 6	2, 3	71,8%
83	0, 2	0, 2	71,6%
84	3, 5	2	71,5%
85	0, 4	2, 4	71,5%
86	3, 4	0, 3	71,3%
87	2, 4	2	71,2%
88	1, 6	1	71,2%
89	1, 5	2	71,1%
90	2, 5	0, 4	71,1%
91	0, 2	2, 3	71,1%
92	3, 5	0	71,1%
93	3, 6	0, 4	71,1%
94	0, 6	0, 4	71,0%
95	3, 6	3, 4	71,0%
96	1	0, 3	70,9%
97	3, 4	0, 2	70,8%
98	3, 5	3	70,7%
99	2, 3	2, 4	70,7%
100	0, 5	3	70,6%
101	0, 6	3, 4	70,6%
102	1, 6	0, 2	70,6%
103	2, 3	0, 2	70,3%
104	1, 2	2, 4	70,2%
105	0, 4	2, 3	70,2%
106	1, 3	2	70,1%
107	2, 4	0, 3	70,1%
108	0, 2	2, 4	70,0%
109	0, 2	0, 3	69,8%
110	1, 5	3, 4	69,5%
111	3, 4	2, 3	69,5%
112	2, 3	2, 3	69,4%
113	1, 3	0, 3	69,3%
114	1, 5	0, 4	69,2%
115	0, 1	0, 3	69,2%
116	1, 6	2, 4	69,1%
117	4, 5	3	68,8%
118	1	0, 4	68,7%
119	4	3	68,5%
120	4, 6	0, 3	68,4%
121	1	3, 4	68,4%
122	2, 3	3, 4	68,3%
123	0, 1	2	68,3%
124	0, 3	0	68,3%
125	0, 3	2	68,2%
126	0, 2	3, 4	68,2%
127	0, 5	4	68,0%
128	2, 3	0, 3	68,0%
129	0, 5	0, 4	67,8%
130	4	0	67,6%
131	5	0	67,4%
132	2, 4	3, 4	67,3%
133	1, 3	3, 4	67,2%
134	4, 6	2	67,2%
135	2, 3	0, 4	67,1%
136	4, 5	0	67,1%
137	0, 2	0, 4	67,0%
138	1, 6	0, 1	66,9%
139	3, 4	3	66,8%
140	2, 6	0, 3	66,6%
141	0, 1	3, 4	66,4%
142	0, 5	0, 3	66,2%
143	0, 3	3	66,2%
144	1, 3	0, 4	66,2%
145	1, 6	1, 4	65,9%
146	1, 6	1, 3	65,8%
147	2, 6	2	65,7%
148	4	0, 4	65,5%
149	0, 1	0, 4	65,5%
150	2, 5	3	65,3%
151	2, 4	0, 4	65,3%
152	0, 5	3, 4	64,9%
153	3, 5	0, 4	64,8%

154	4, 6	3, 4	64,8%
155	0, 4	3	64,7%
156	1, 4	2	64,5%
157	5	4	64,5%
158	2	3	64,4%
159	1, 6	1, 2	64,3%
160	6	3	64,2%
161	2, 5	0	64,1%
162	4	0, 3	63,9%
163	3, 4	0	63,9%
164	2	0, 4	63,8%
165	0, 5	0, 2	63,7%
166	1, 4	0, 3	63,6%
167	1, 2	2	63,6%
168	5	3	63,5%
169	0, 3	4	63,3%
170	1, 2	0, 3	63,3%
171	2	0	63,2%
172	2	0, 3	63,2%
173	3, 5	0, 3	63,0%
174	4	3, 4	63,0%
175	6	0	62,9%
176	0, 4	0	62,9%
177	0, 5	2, 4	62,8%
178	2, 6	3, 4	62,8%
179	2	3, 4	62,8%
180	4	4	62,3%
181	4, 6	0, 4	62,1%
182	2, 3	3	62,1%
183	1, 4	1	62,0%
184	3, 5	3, 4	61,9%
185	2	0, 2	61,9%
186	1, 2	1	61,8%
187	0, 5	2, 3	61,6%
188	5	2	61,4%
189	2, 6	0, 4	61,4%
190	1, 6	2	61,4%
191	2, 3	0	61,3%
192	0, 3	0, 4	61,3%
193	5, 6	3	61,2%
194	0, 2	3	60,9%
195	0	0	60,7%
196	1, 4	3, 4	60,7%
197	3, 5	0, 2	60,6%
198	1, 6	0, 3	60,4%
199	5, 6	0	60,3%
200	4	0, 2	60,1%
201	0, 2	0	60,1%
202	1	3	60,1%
203	0, 3	0, 3	60,1%
204	3, 5	2, 3	60,0%
205	3, 6	3	59,9%
206	2	2, 4	59,7%
207	3	0	59,6%
208	0, 6	3	59,4%
209	1, 4	0, 4	59,4%
210	2	2, 3	59,2%
211	1, 2	3, 4	59,1%
212	2, 5	4	59,1%
213	1	0	59,0%
214	1, 2	0, 4	58,8%
215	4, 5	4	58,7%
216	3, 5	2, 4	58,5%
217	3, 6	0	58,4%
218	0, 3	3, 4	58,3%
219	5		58,3%
220	0, 3	0, 2	58,2%
221	4	2, 3	58,0%
222	1, 6	3, 4	57,9%
223	1, 5	3	57,9%
224	0, 6	0	57,6%
225	1, 6	0, 4	57,5%
226	1, 5	0	57,5%
227	6	4	57,4%
228	0	3	57,2%
229	0	2	57,0%
230	4	2, 4	56,9%
231	1, 5	1	56,9%
232	3	3	56,8%
233	0, 3	2, 4	56,5%
234	0	4	56,4%
235	4, 6	1	56,4%

236	0, 3	2, 3	56,3%
237	1, 3	0	56,2%
238	1, 4	1, 3	56,2%
239	1, 3	3	56,0%
240	3	4	56,0%
241	3, 4	4	55,9%
242	1, 4	0, 1	55,9%
243	2	4	55,7%
244	2, 6	1	55,7%
245	5, 6	4	55,4%
246	0, 1	1	55,3%
247	1, 2	0, 1	54,8%
248	1, 4	1, 4	54,7%
249	0, 4	4	54,7%
250	0, 6	4	54,7%
251	1	4	54,6%
252	1, 2	1, 3	54,5%
253	0, 1	3	54,5%
254	0, 1	0	54,3%
255	1, 2	1, 4	54,3%
256	1, 4	1, 2	54,2%
257	2, 3	4	54,0%
258	3, 6	4	53,8%
259	1, 5	4	53,8%
260	2, 4	3	53,8%
261	0, 2	4	53,7%
262	1, 3	1	53,3%
263	1, 2	1, 2	53,2%
264	3	2	52,9%
265	2, 4	0	52,8%
266	5	0, 3	52,7%
267	1, 3	4	52,0%
268	0, 1	4	51,9%
269	1, 5	0, 1	51,2%
270	1, 5	1, 3	51,1%
271	5, 6	1	50,7%
272	7		50,4%
273	1, 5	1, 4	50,0%
274	2, 6	0	49,9%
275	2, 6	3	49,9%
276	3		49,7%
277	5	0, 4	49,5%
278	4, 6	1, 3	49,5%
279		5	49,5%
280	1, 5	1, 2	49,3%
281	4, 6	3	49,3%
282	4, 6	0, 1	49,2%
283	1, 2	0	49,0%
284	0, 6	1	48,8%
285	0, 1	1, 3	48,7%
286	0, 1	0, 1	48,6%
287	5	2, 3	48,5%
288	5	0, 2	48,3%
289	5	3, 4	48,2%
290	0, 1	1, 2	48,1%
291	0, 1	1, 4	48,0%
292	2, 6	0, 1	48,0%
293	1, 4	0	47,8%
294	4, 6	1, 2	47,8%
295	4, 6	1, 4	47,7%
296	0	0, 3	47,7%
297	1, 3	0, 1	47,5%
298	1, 6	3	47,4%
299	1, 2	3	47,4%
300	4, 6	0	47,4%
301	2, 6	1, 3	47,2%
302	0	0, 4	47,1%
303	1	1	47,1%
304	1, 3	1, 3	47,1%
305	1, 6	0	46,8%
306	1, 3	1, 4	46,7%
307	2, 6	4	46,7%
308	1, 4	3	46,5%
309	3, 6	1	46,3%
310	1, 6	4	46,2%
311	1, 3	1, 2	46,2%
312	0	2, 4	46,2%
313	0	0, 2	46,1%
314	0	3, 4	46,1%
315	2, 6	1, 4	46,0%
316	5, 6	1, 3	45,9%
317	0	2, 3	45,8%

318	2, 6	1, 2	45,6%
319	1, 2	4	45,5%
320	5, 6	0, 1	45,5%
321	3	0, 3	45,5%
322	2, 4	4	45,2%
323	4, 6	4	45,0%
324	5, 6	1, 2	44,8%
325	5	2, 4	44,1%
326	1, 4	4	43,9%
327	3	0, 4	43,9%
328	3	0, 2	43,6%
329	0, 6	1, 3	43,6%
330	5, 6	1, 4	43,4%
331	0, 6	1, 2	42,8%
332	0, 6	0, 1	42,6%
333	1	0, 1	42,3%
334	4, 5	1	42,2%
335	0, 6	1, 4	42,0%
336	3	2, 3	41,9%
337	3, 6	1, 3	41,9%
338	3, 6	0, 1	41,9%
339	1	1, 3	41,7%
340	1	1, 2	41,7%
341	2, 4	1	41,7%
342	3, 6	1, 2	41,2%
343	1	1, 4	41,2%
344	6	1	40,7%
345	3	3, 4	39,9%
346	2		39,8%
347	6		39,5%
348	3, 6	1, 4	39,5%
349	4, 5	1, 3	39,4%
350		3	39,1%
351	1		39,0%
352	4, 5	0, 1	39,0%
353	4, 5	1, 2	38,8%
354	2, 5	1	38,8%
355	3	2, 4	38,6%
356	2, 4	0, 1	38,5%
357	3, 4	1	38,5%
358	0, 4	1	38,4%
359	2, 4	1, 3	38,3%
360	4		38,2%
361	6	1, 3	38,2%
362	6	0, 1	37,7%
363	6	1, 2	37,5%
364	2, 5	1, 3	37,3%
365	2, 4	1, 2	37,2%
366	3, 5	1	37,0%
367	0, 2	1	36,9%
368	0, 4	1, 3	36,9%
369	4, 5	1, 4	36,7%
370	0, 5	1	36,7%
371		2	36,7%
372	2, 5	0, 1	36,7%
373	2, 3	1	36,6%
374	3, 4	0, 1	36,6%
375	3, 4	1, 3	36,5%
376	3, 4	1, 2	36,5%
377	2, 5	1, 2	36,4%
378	3, 5	1, 3	36,1%
379	3, 5	0, 1	36,1%
380	0, 4	1, 2	36,1%
381	0, 4	0, 1	36,1%
382	6	1, 4	36,0%
383	3, 5	1, 2	35,8%
384	2, 4	1, 4	35,8%
385	0, 5	0, 1	35,7%
386	0, 5	1, 3	35,6%
387	0, 5	1, 2	35,5%
388	0, 4	1, 4	35,3%
389	0, 2	1, 3	35,2%
390	0, 2	0, 1	35,2%
391	0, 2	1, 2	35,2%
392	2, 5	1, 4	35,0%
393	2, 3	0, 1	35,0%
394	2, 3	1, 3	34,9%
395	0, 3	1	34,8%
396	3, 5	1, 4	34,8%
397	0, 2	1, 4	34,7%
398	4	1, 2	34,6%
399	4	1, 3	34,6%

400	4	0, 1	34,5%
401	2, 3	1, 2	34,5%
402	4	1	34,5%
403	5	0, 1	34,5%
404	3, 4	1, 4	34,5%
405	5	1	34,4%
406	0, 3	0, 1	34,4%
407	5	1, 3	34,3%
408	0, 3	1, 3	34,3%
409	0, 5	1, 4	34,3%
410	0, 3	1, 2	34,3%
411	5	1, 2	34,3%
412	2, 3	1, 4	34,1%
413	2	1	34,0%
414	0	1, 2	34,0%
415	0	1	34,0%
416	2	0, 1	33,9%
417	0, 3	1, 4	33,9%
418	5	1, 4	33,9%
419	2	1, 2	33,9%
420	0	1, 3	33,8%
421	0	0, 1	33,8%
422	4	1, 4	33,8%
423	3	1, 2	33,8%
424	3	1	33,7%
425	3	1, 3	33,7%
426	2	1, 3	33,7%
427	0	1, 4	33,7%
428	2	1, 4	33,6%
429	3	0, 1	33,6%
430	3	1, 4	33,6%
431		1	33,5%
432		4	33,5%

10 References

Aggarwal, C. C. and Reddy, C. K. (2013) *Data Clustering: Algorithms and Applications*. Taylor \& Francis (Chapman \& Hall/CRC Data Mining and Knowledge Discovery Series). Available at: https://books.google.se/books?id=edl7AAAAQBAJ.

Aquino, F. *et al.* (2017) 'The web and public confidence in MMR vaccination in Italy', *Vaccine*. Elsevier, 35(35), pp. 4494–4498.

Balmas, M. (2014) 'When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism', *Communication research*. Sage Publications Sage CA: Los Angeles, CA, 41(3), pp. 430–454.

Cantarella, M., Fraccaroli, N. and Volpe, R. (2020) 'Does fake news affect voting behaviour?' CEIS Working Paper.

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

Dubé, E., Vivion, M. and MacDonald, N. E. (2015) 'Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications', *Expert review of vaccines*. Taylor \& Francis, 14(1), pp. 99–117.

Gottfried, J. and Shearer, E. (2019) 'News use across social media platforms 2016'. Pew Research Center.

Horne, B. and Adali, S. (2017) 'This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news', in *Proceedings of the International AAAI Conference on Web and Social Media*.

Jolley, D. and Douglas, K. M. (2014) 'The effects of anti-vaccine conspiracy theories on vaccination intentions', *PloS one*. Public Library of Science San Francisco, USA, 9(2), p. e89177.

Jurafsky, D. and Martin, J. H. (2021) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Third edit. Prentice hall Upper Saddle River, NJ. Available at: https://web.stanford.edu/~jurafsky/slp3/.

Ketchen, D. J. and Shook, C. L. (1996) 'The application of cluster analysis in strategic management research: an analysis and critique', *Strategic management journal*. Wiley Online Library, 17(6), pp. 441–458.

Lazer, D. M. J. *et al.* (2018) 'The science of fake news', *Science*. American Association for the Advancement of Science, 359(6380), pp. 1094–1096.

Levi, O. *et al.* (2019) 'Identifying nuances in fake news vs. satire: Using semantic and linguistic cues', *arXiv preprint arXiv:1910.01160*.

Moravec, P., Minas, R. and Dennis, A. R. (2018) 'Fake news on social media: People believe what they want to believe when it makes no sense at all', *Kelley School of Business Research Paper*, (18–87).

Pilehvar, M. T. and Camacho-Collados, J. (2020) *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning*. Morgan \& Claypool Publishers (Synthesis Lectures on Human Language Technologies). Available at: https://books.google.se/books?id=U90MEAAAQBAJ.

Robertson, S. (2004) ‘Understanding inverse document frequency: on theoretical arguments for IDF’, *Journal of documentation*. Emerald Group Publishing Limited.

scikit-learn developers (2021a) *1.9. Naive Bayes*. Available at: https://scikit-learn.org/stable/modules/naive_bayes.html.

scikit-learn developers (2021b) *TfidfTransformer*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html.

Silverman, C. and Singer-Vine, J. (2016) ‘Most Americans who see fake news believe it, new survey says’, *BuzzFeed news*, 6(2).

Stone, J. V (2013) ‘Bayes’ rule: a tutorial introduction to Bayesian analysis’. Sebtel Press.

Verma, P. K. *et al.* (2021) ‘WELFake: Word Embedding Over Linguistic Features for Fake News Detection’, *IEEE Transactions on Computational Social Systems*. IEEE.