# Analysis of Differences in Texts between Real and Fake News

Cristian Lopez, Karina Sindermann, Quan Nguyen, Sebastian Vazquez-Gasty, Jackson Maschman
University of Nebraska - Lincoln

## 1 INTRODUCTION

Social media has undoubtedly increased our online presence and connected us with like-minded people from around the world. It has also allowed us to foster relationships and share memorable moments with our loved ones. However, the rise of fake news on social media has presented challenges for companies seeking to control its spread while maintaining freedom of speech. The issue has become even more significant with the manipulation of these platforms by companies or individuals to spread false information. While many companies have attempted to strike down these contents, the damage has already been done to individuals or entities. To prevent further harm, many companies have started developing preventative systems to identify and prevent the spread of fake news. One such technique is to employ a team of reviewers to examine each post or content and determine its reliability[1]. However, with the sheer amount of data being generated daily, this method is impractical, and companies need a more efficient approach. Our team aims to explore ways in which companies can analyze billions of bytes of data effectively. Specifically, we will examine differences in word choice, grammar, and sentence structure between real and fake news to identify patterns that can help companies narrow down the list of content needing review. This approach will enable companies to allocate their limited human resources to the most likely-to-be-false news.

## 2 OBJECTIVES

1) Data preprocessing. Tokenizing texts in the inputs.
2) Do association mining between word choices and fake news.
3) Research existing models that can produce text patterns for fake news and real news and obtain their patterns. Compare these patterns to patterns generated from our association mining.
4) Evaluate the accuracy and efficiency of association mining in objective 2 and explore alternative methods to identify fake news text patterns.

## 3 RELATED WORK

Fake news detection is defined as a binary classification problem [2]. The fake news detection problem is a data mining problem because it consists of two stages: (a) feature extraction and (b) model building. The feature extraction phase seeks to formalize the mathematical structure of news content and related auxiliary data. The model construction phase further develops machine learning models to more accurately distinguish between fake and legitimate news based on the feature representations. In literature, many approaches have been proposed to potentially identify fake news. For instance:

i. Visual-based: by using a classification framework, fake pictures were detected based on a variety of user preferences and features in social media that are manually designed with intuitive mode [3]. Also, visual statistical features have been implemented to identify fake news [4].
ii. Linguistics-based: main feature categories such as lexical, grammatical, and syntactic, are used in qualitative and quantitative analysis to capture fake news [5].
iii. Post-based: fake news based on personal opinions. In coordination with linguistics-based approaches, it is intended to identify support or disavowal of certain news, and thus, the reliability of posts is evaluated by their credibility features [6]
iv. Network-based: the key idea is to track the source and the spread of misinformation based on networks of users that share common interests [7].

In this project, we would like to identify fake news focused on linguistics-based features by using text mining association and comparing our results to state-of-the-art models to provide robust criteria for identifying unreliable news from any social media.

## 4 PROBLEM DEFINITION

In recent years, the issue of fake news has become a major concern across the globe. Fake news refers to deliberately fabricated or manipulated information that is spread to mislead people or create a false narrative. This type of misinformation can be spread through various media channels, such as social media, online news platforms, and traditional media outlets. Fake news can cause harm to individuals, organizations, and even society as a whole, by influencing opinions, beliefs, and behaviors, and can lead to negative consequences, such as political unrest, social tensions, and even violence. As a result, there is a growing need for effective methods to detect and prevent the spread of fake news.

In order to analyze fake news, text mining is often employed. Text mining is the process of finding meaningful patterns and new insights from unstructured texts. Text mining techniques can be used to analyze language patterns in news articles and social media posts to identify potential instances of fake news. This project aims to apply various techniques to analyze the features distinguishing real news from fake news. The inputs for this scope are primarily social media sentences, among other attributes and the output will be how likely new sentences are to be fake. It is important to mention that for this project, we only work with English-language texts.

## 5 DATASET

We use the popular Fake News Datasets LIAR [9] and part of the FakeNewsNet [8] Dataset. Both Datasets were constructed using the fact-checking website PolitiFact. PolitiFact is an independent, nonpartisan online fact-checking website. It primarily rates the accuracy of claims or statements made by political news and politicians in the U.S.

GossipCop also fact-checked part of the news in the FakeNewsNet Dataset. GossipCop (now suggest) checks fake news from the entertainment and celebrity sectors in the U.S. published in magazines and web portals.

LIAR [9] is a publicly available Dataset for fake news detection. It contains 12836 short statements from 2007 to 2016. A POLITIFACT editor evaluated each statement. For the Truthfulness ratings, six finely nuanced labels are used (true, mostly true, half-true, mostly false, false, and pants on fire). Furthermore, the Dataset contains the statement, its subject, the context of when it was made, the speaker's name, job title, home state, and party affiliation, and their historical count of inaccurate statements. An example entry of this database is shown in Tab. 1.

Although the authors state that the database consists of 12836 instances, we can only access 12791 values (see Tab. 2).

| | |
|---|---|
| label | 4 (barely true) |
| statement | "Most of the (Affordable Care Act) has already in some sense been waived or otherwise suspended." |
| subject | "Healthcare" |
| speaker | "George Will" |
| job_title | "Columnist" |
| state_info | "Maryland" |
| party_affiliation | "columnist" |
| barely_true_counts | 7 |
| false_counts | 6 |
| half_true_counts | 3 |
| mostly_true_counts | 5 |
| pants_on_fire_counts | 1 |
| context | "Comments on 'Fox News Sunday'" |

**Tab. 1:** Example of LIAR data set.

| label | train | test | validation | Sum |
|---|---|---|---|---|
| barely true | 1654 | 212 | 237 | 2103 |
| FALSE | 1995 | 249 | 263 | 2507 |
| half-true | 2114 | 265 | 248 | 2627 |
| mostly true | 1962 | 241 | 251 | 2454 |
| pants-fire | 839 | 92 | 116 | 1047 |
| TRUE | 1676 | 208 | 169 | 2053 |
| Sum | 10240 | 1267 | 1284 | 12791 |

**Tab. 2:** Description of the LIAR data set

The FakeNewsNet [8] Dataset for detecting fake news provides a comprehensive social media context. The Dataset includes articles and associated tweets checked by PolitiFact or GossipCop. It contains 467 thousand tweets in the PolitiFact Dataset and 1.25 million tweets in the GossipCop Dataset, labeling them as either true or fake.

Due to Twitter's privacy policy and the copyrights of news publishers, the full Dataset cannot be made public. We will use a small sample of the Dataset, which is publicly available. This sample consists of 4 files that differ by the truth value of the story (false or real) and the verifier (PolitiFact or GossipCop). The sample contains 23424 verified news titles (see Tab. 3). The attributes of the Datasets are id, URL of the news, title, and Twitter id.

| | Real | Fake | Sum |
|---|---|---|---|
| GossipCop | 16818 | 5335 | 22153 |
| PolitiFact | 798 | 473 | 1271 |
| Sum | 17616 | 5808 | 23424 |

**Tab. 3:** Description of the FakeNewsNet data set

# 6 APPROACH

### 6. 1. Preprocessing

The first step in our approach is data preprocessing. We started with two datasets: the Liar dataset and the FakeNewsNet dataset. The Liar dataset contains statements labeled with six possible truth ratings: "pants-fire," "false," "barely-true," "half-true," "mostly-true," and "true," while the FakeNewsNet dataset consists of news articles labeled as either true or false.

Once the datasets were loaded, we combined them into a single dataset for ease of processing. We removed rows with missing values and converted labels to numerical values. In the Liar dataset, we mapped the labels "pants-fire," "false," and "barely-true" to 0 and "half-true," "mostly-true," and "true" to 1. Similarly, in the FakeNewsNet dataset, we mapped the label "false" to 0 and "true" to 1.

We defined a function to preprocess the text data by removing punctuation, converting the text to lowercase, and removing stop words. We applied this function to the "statement" column of the combined Dataset.

Once the text data had been cleaned, we converted the statements into a matrix of numerical features using the TF-IDF (term frequency inverse document frequency) Vectorizer.

TF-IDF stands for term frequency-inverse document frequency and consists of term frequency (TF) and inverse document frequency (IDF). The former measures the number of times a word appears in a document, divided by the total number of words, while the latter measures how much information a word provides across the entire collection of documents. TF-IDF assigns a higher weight to words specific to a particular document and a lower weight to words common across all documents in the collection.

In summary, preprocessing was to prepare the data for analysis and ensure that it was in a usable format. It involved cleaning and converting the data into a numerical format.

### 6. 2. Apriori Algorithm

The Apriori algorithm is an association rule mining algorithm that identifies frequent patterns or associations among different features. It can be used to identify common patterns in the language used in fake news articles, such as the frequent occurrence of certain sensationalist words or phrases. By doing so, it is possible to identify patterns in the sources or websites that publish fake news articles or in the behavior of social media accounts that spread fake news, such as frequently sharing articles from known fake news sources. The use of Apriori can help find patterns or characteristics that are indicative of fake or misleading information. Once these patterns have been identified, they can be used in classification algorithms to help identify potential fake news.

For the Apriori algorithm, we need to identify support and confidence thresholds. Using inappropriate thresholds can lead to either no rules or too many rules. To avoid guessing which values may seem reasonable, we obtained the 20 most words and their frequencies, as shown in Figure 1.

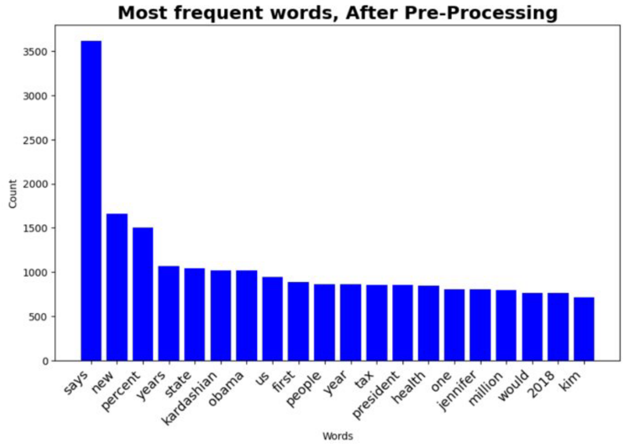


**Fig. 1:** 20 most frequent words

We realized that "Kardashian" and "Kim" are among the most frequent words. Then, we calculated the support of this itemset of these two words as

$$s = \frac{507}{35987} = 0.014\ (1.4\%)$$

Thus, based on this result, we are guided to set the minimum support values as they are going to be presented in the following.

To be able to find such associations/rules, the following steps are needed.

Step 1. Load and preprocess sentences. After the preprocessing, we decided to perform three experiments. They consist in using a) a whole data set, b) real news, and fake news. For any of the three cases, we perform the following three steps.

Step 2. Convert sentences to transactions. The sentences are converted into a list of transactions. A transaction is a list of items (or words, in this case) associated with each other. To do this, we iterate over each sentence in the list of sentences and split it into individual items (words). We then append each transaction to a list of transactions.

Step 3: Create transaction Dataset. A TransactionEncoder object is created, which converts the list of transactions into a one-hot encoded array, where each column represents a unique item in the transactions and each row represents a single transaction. The one-hot encoded array is then converted to a suitable format, Pandas DataFrame in our case, for further analysis.

Step 4: Apply the Apriori algorithm. In this step, we generate a list of frequent itemsets from the transaction Dataset using the minimum support threshold for an itemset to be considered frequent. And then extract association rules from the frequent itemsets.

### 6.3. Ensemble Method using Random Forest and Gradient Boosting

After conducting an analysis using the Apriori algorithm, we explored state-of-the-art methods to gain further insights. Our research led us to a paper titled "Text-mining-based Fake News Detection Using Ensemble Methods" [11], highlighting Random Forest and Gradient Boosting as the most effective ensemble methods for detecting fake news.

Both Random Forest and Gradient Boosting are ensemble methods, but they differ in their approach. Random Forest employs a bagging technique, where multiple decision trees work independently. On the other hand, Gradient Boosting utilizes a boosting technique where multiple decision trees collaborate iteratively to improve results.

Initially, we used the default parameters for both methods. However, since our Dataset exhibited class imbalance, we addressed this issue by optimizing the classifiers through stratified cross-validation. We experimented with various parameter combinations while maintaining the data proportion between training and testing sets.

To evaluate the results, we examined the confusion matrix and employed different evaluation metrics such as accuracy and F-1 score. We compared the performance between the initial run and the optimized run to assess the effectiveness of our approach.

**6. 4. Qualitative Data Analysis**

Finally, we conducted an additional analysis using linguistic features and qualitative data analysis for our last method. For this method, we looked at each sentence's usage of pronouns, adjectives, and sensational words, as well as the overall sentiment of it, to discern whether any features were more likely to result in false news.

Therefore, for this step, each word in each sentence was converted to its word form and roles. For example, John is considered a noun and a person, while "most beautiful" is considered an adjective and superlative. Because our Dataset contained 65% of real news, we weighed the results to remove the impact of the initial distribution.

While running existing models in Quirkos, we also removed some linguistic features that seemed irrelevant to the research while telling the model to focus on common features for fake news patterns such as tone, tense, pronoun, adjectives, and words carrying emotional connotations.

For the results, we looked at the percentage of fake and truthful news in each linguistic feature.

## 7 Results

### 7.1 Apriori Algorithm

7.1.1. The entire dataset

We used the whole data set (35987 sentences) with a support value of 0.4% (143 words) and a confidence value of 90%. We obtained 482 frequent items and 32 rules. Fig. 2 illustrated some results of the rules sorted by confidence level.

| | antecedents | consequents | support | confidence |
|---|---|---|---|---|
| 12 | (york) | (new) | 0.004585 | 1.000000 |
| 21 | (justin, gomez) | (selena) | 0.004807 | 1.000000 |
| 4 | (pitt) | (brad) | 0.010976 | 0.994962 |
| 22 | (markle, harry) | (meghan) | 0.005391 | 0.994872 |
| 30 | (markle, harry, prince) | (meghan) | 0.005252 | 0.994737 |
| 17 | (angelina, pitt) | (brad) | 0.005196 | 0.994681 |
| 8 | (stefani) | (gwen) | 0.004752 | 0.994186 |
| 20 | (jolie, pitt) | (brad) | 0.004224 | 0.993464 |
| 18 | (jolie, pitt) | (angelina) | 0.004224 | 0.993464 |
| 28 | (jolie, angelina, pitt) | (brad) | 0.004196 | 0.993421 |

**Fig. 2:** Rules obtained from the entire dataset sorted based on the confidence level. Support level = 0.4%

To uncover additional rules, we experimented with confidence values of 0.5 and 0.1. However, the results yielded only names and names combined with the word "says" without revealing any noteworthy rules.

For a subsequent attempt, we adjusted the support threshold to 0.2% (equivalent to 71 words) while maintaining a confidence level of 90%. This yielded 1,225 frequent items and 137 rules. Figure 3 displays these rules in descending order based on their confidence scores.

| | antecedents | consequents | support | confidence |
|---|---|---|---|---|
| 36 | (york) | (new) | 0.004585 | 1.0 |
| 126 | (foxx, holmes, katie) | (jamie) | 0.002418 | 1.0 |
| 30 | (theroux) | (justin) | 0.003696 | 1.0 |
| 125 | (foxx, jamie, katie) | (holmes) | 0.002418 | 1.0 |
| 32 | (kidman) | (nicole) | 0.002890 | 1.0 |
| 88 | (holmes, jamie) | (katie) | 0.002862 | 1.0 |
| 72 | (foxx, holmes) | (jamie) | 0.002418 | 1.0 |
| 124 | (holmes, foxx, jamie) | (katie) | 0.002418 | 1.0 |
| 42 | (lopez, alex) | (jennifer) | 0.002445 | 1.0 |
| 119 | (stefani, shelton, blake) | (gwen) | 0.003390 | 1.0 |
| 45 | (rodriguez, lopez) | (alex) | 0.002223 | 1.0 |

**Fig. 3:** Rules obtained from the entire dataset sorted based on the confidence level. Support level = 0.2%

Since we don't find promising rules, we attempt to find rules that may guide us in detecting fake news patterns, we decided to split the data and work separately with the real and fake news.

7.1.2 Real news dataset

There are 24,575 sentences in this group, and we set the support value as 0.4% (equivalent to 98 words). With this setting, we were able to obtain 421 frequent items and 61 frequent rules. Similar to the previous case, the results are shown in Fig. 4.

| Confidence | # of Rules | Interesting rules |
|---|---|---|
| 0.9 | 63 | 1. (wedding, harry , meghan) =>markle |
| 0.5 | 218 | 1. (wedding, harry , meghan) =>markle<br>2. (wedding, harry) => Meghan<br>3. (red , carpet) => award<br>4. (people, choice) =>award |
| 0.1 | 516 | 1. (wedding, harry) => Meghan<br>2. (red , carpet) => award<br>3. (people, choice) =>award<br>4. job => lost<br>5. job => created |

**Tab. 4:** Results after varying confidence value for Fake News dataset

| | antecedents | consequents | support | confidence |
|---|---|---|---|---|
| 4 | (york) | (new) | 0.005005 | 1.000000 |
| 7 | (harry, markle) | (meghan) | 0.004720 | 1.000000 |
| 12 | (harry, markle, prince) | (meghan) | 0.004639 | 1.000000 |
| 0 | (carpet) | (red) | 0.008057 | 0.994975 |
| 1 | (gomez) | (selena) | 0.006307 | 0.993590 |
| 10 | (markle, prince) | (meghan) | 0.005249 | 0.992308 |
| 3 | (markle) | (meghan) | 0.009237 | 0.991266 |
| 8 | (harry, markle) | (prince) | 0.004639 | 0.982759 |
| 11 | (harry, markle, meghan) | (prince) | 0.004639 | 0.982759 |
| 13 | (harry, markle) | (meghan, prince) | 0.004639 | 0.982759 |
| 6 | (swift) | (taylor) | 0.006633 | 0.964497 |
| 2 | (housewives) | (real) | 0.004924 | 0.952756 |
| 5 | (united) | (states) | 0.009766 | 0.952381 |
| 9 | (harry, meghan) | (prince) | 0.005738 | 0.940000 |

**Fig. 4:** Rules obtained from real news dataset sorted based on the confidence level.

We also tried with support as 0.002 and we got 1117 frequent items and then, we changed the confidence. Results are presented in Tab. 4.

7.1.3. Fake news

In this subset, there were a total of 11,412 sentences. To determine the relevance of the data, a support value of 0.4% (equivalent to 45 words) was established. Additionally, a confidence threshold of 90% was set. After executing the algorithm, we identified 610 frequent items and 138 frequent rules. These results are presented in Figure 5.

| | antecedents | consequents | support | confidence |
|---|---|---|---|---|
| 39 | (affleck, garner) | (jennifer) | 0.004118 | 1.0 |
| 118 | (bieber, gomez) | (selena, justin) | 0.008675 | 1.0 |
| 116 | (bieber, selena, gomez) | (justin) | 0.008675 | 1.0 |
| 35 | (affleck, garner) | (ben) | 0.004118 | 1.0 |
| 111 | (jennifer, aniston, theroux) | (justin) | 0.005608 | 1.0 |
| 37 | (affleck, jennifer) | (ben) | 0.004819 | 1.0 |
| 107 | (pitt, jennifer, aniston) | (brad) | 0.008237 | 1.0 |
| 40 | (lopez, alex) | (jennifer) | 0.004206 | 1.0 |
| 43 | (angelina, divorce) | (brad) | 0.004381 | 1.0 |
| 102 | (affleck, garner) | (jennifer, ben) | 0.004118 | 1.0 |
| 49 | (pitt, aniston) | (brad) | 0.008587 | 1.0 |
| 100 | (affleck, ben, garner) | (jennifer) | 0.004118 | 1.0 |

**Fig. 5:** Fake news. Rules sorted based on confidence.

Among 138 rules, these two rules stood out: “(Angelina, divorce) => brad” and “(law, health) => care”

Similar to the real news dataset experiment, we used a support threshold of 0.2% (42 words) and got 1576 frequent items. Then, by keeping the same confidence, the number of rules is 484. However, no interesting results were found.

Finally, we explored both real and fake news by using support and confidence as 0.4% and 50%, respectively. Although the rules may be weak, we found that both cases shared three categories:

1. Celebrities’ names: Harry, Selena, Kardashian, Meghan, Kylie, etc.
2. Entertainment, music, and awards: choice, carpet housewives, wedding, red, health, awards, etc.
3. Political figures: Donald, Hillary, Clinton, Trump, Obama, Barack

Interestingly, for fake news, we found other words related to:

1. Legal issues: administration, illegal, supreme, report, court, divorce, insurance,
2. Immigration: immigrants, social
3. Government: gov, administration.

In the case of both real and fake news, celebrities, entertainment, music, and political figures are some of the main topics covered. This is not surprising as these are all popular and newsworthy topics that are likely to generate a lot of interest and attention from readers and viewers. The categories commonly associated with fake news, such as legal issues, immigration, government, and political figures, suggest a focus on controversial and divisive issues likely to generate strong emotions and opinions among readers. This may be a deliberate strategy to attract attention and generate clicks or shares on social media platforms.

**7.2 Ensemble Methods**

We first investigate the confusion matrix that we got from using the classifiers with the default parameters.

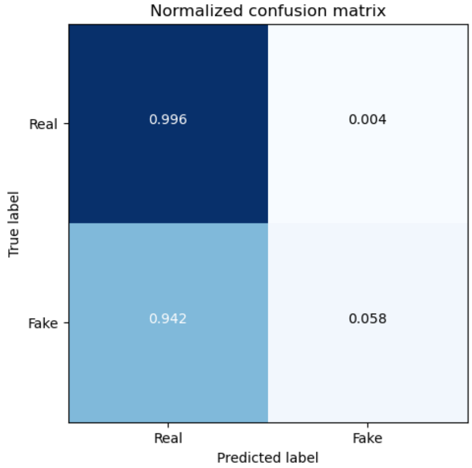


**Fig. 6:** Confusion matrix: Random Forest

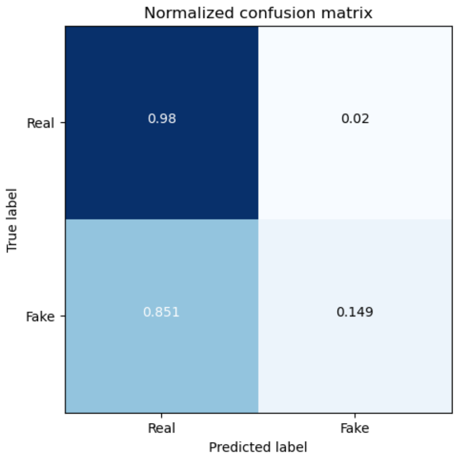


**Fig. 7:** Confusion matrix: Gradient Boosting

Upon analyzing the results, we observed a low rate of true negatives and a high rate of false positives. This discrepancy indicated that our classifiers required improvement. To address this issue, we implemented a procedure called GridsearchCV.

GridsearchCV allowed us to define a set of parameters to be tested. The method then systematically explored various combinations of these parameters and applied stratified cross-validation with each classifier. By doing so, we aimed to find the optimal parameter configuration that would enhance the performance of the classifiers.

| Classifier | RF | GB | RF Opt | GB Opt |
|---|---|---|---|---|
| Accuracy | 0.69 | 0.71 | 0.75 | 0.74 |
| Precision | 0.78 | 0.74 | 0.73 | 0.71 |
| Recall | 0.53 | 0.56 | 0.66 | 0.64 |
| F1-Score | 0.46 | 0.54 | 0.67 | 0.65 |
| Specificity | 0.06 | 0.15 | 0.4 | 0.37 |

**Tab. 5:** Scores in each classifier

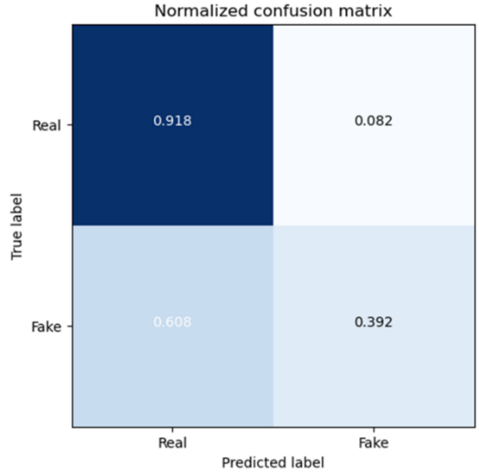


**Fig. 8:** Confusion matrix: Random Forest optimized.

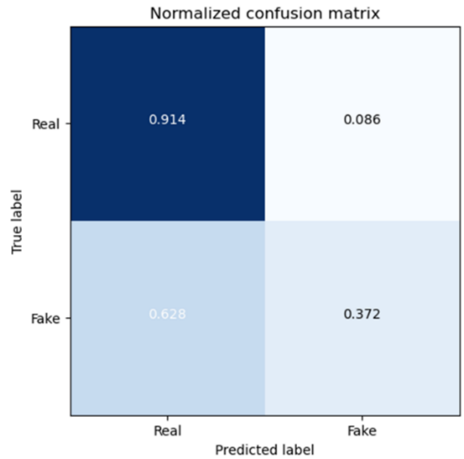


**Fig. 9:** Confusion matrix: Gradient Boosting optimized.

The results indicate a noticeable improvement in the performance of the classifiers. There is an increase in true negatives and a decrease in false positives, as observed in the data. Table 5 further demonstrates an increase in specificity while maintaining a consistent score for sensitivity or precision. This suggests that the classifiers' ability to correctly identify true negatives and true positives has been enhanced without compromising the overall model.

Although the classifiers may not have exhibited exceptional performance, due to the high class imbalance in the data, it is evident that the GridsearchCV method is an effective procedure for testing different parameter combinations and selecting the most suitable ones. This approach proves to be valuable for optimizing classifier performance and addressing the challenges posed by imbalanced Datasets.

### 7.3. Qualitative Data Analysis

Due to the limitations of the Apriori analysis, we sought to explore other avenues to identify patterns in fake news. As a result, we utilized a linguistic feature analysis approach using Quirkos, a commercial Qualitative Data Analysis (QDA) tool. This tool employed advanced Machine Learning models such as Part-of-speech tagging, Named Entity Recognition, and sentiment analysis to identify linguistic features and produce key patterns among a set of sentences. We narrowed our focus to commonly observed patterns that people use to discern fake news in their daily lives, such as the use of extreme language, sensational headlines, and emotional language. Given the uneven distribution of our Dataset (with 68% consisting of real news), we applied a weighting/scaling approach to ensure unbiased summarization of the data. Our findings highlight the usefulness of linguistic feature analysis in identifying patterns that are not necessarily captured through traditional statistical methods like Apriori.

| Categories | Fake News | Real News |
|---|---|---|
| Superlatives | 43.65% | 56.35% |
| Comparatives | 52.47% | 47.53% |
| Sensational words | 49.36% | 50.64% |
| Pronoun | 50.20% | 49.80% |
| first- and second-person pronoun | 49.80% | 50.20% |
| Positive sentiment | 40.27% | 59.73% |
| Negative Sentiment | 53.17% | 46.83% |

**Tab. 6:** Qualitative Data Analysis Results

| | Positive sentiment | Negative sentiment |
|---|---|---|
| Fake News | 9.2% | 11.6% |
| Real News | 13.7% | 10.3% |

**Tab. 7:** Proportion of positive and negative sentiment news in fake and real news categories respectively.

The analysis of the metrics "comparatives," "sensational words," "pronouns," and "first- and second-person pronoun" did not show a statistically significant difference between real news and fake news, indicating that these features were not useful for detecting fake news in this Dataset. However, superlatives were found to be more common in real news, which was not in line with previous research that has shown how superlatives are often used to exaggerate and sensationalize news stories.

In contrast, the sentiment analysis results revealed that almost 60% of the positive news articles were truthful. This finding led us to examine the distribution of news types in real and fake news respectively (as shown in Tab. 6). Interestingly, we observed a significant difference in the percentage of positive news articles between real and fake news, suggesting that fake news writers were less likely to focus on uplifting and joyful stories, and instead may have been more focused on negative and sensational news. Next, we examined the distribution of sentiment between real and false news.

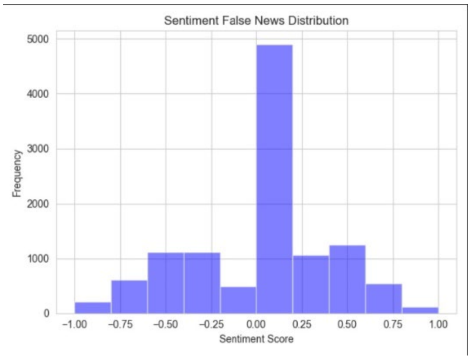


**Fig. 10:** Sentiment False News Distribution

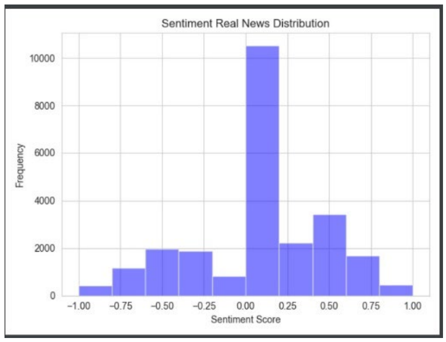


**Fig. 11:** Sentiment Real News Distribution

One noticeable trait was the bimodal nature of both distributions. A bimodal distribution in a histogram of sentiment scores might infer that the Dataset consists of two distinct groups or clusters of data points with different sentiment scores. For example, in the context of sentiment analysis of product reviews, a bimodal distribution may suggest two types of customers: one highly satisfied with the product and another highly dissatisfied group. In our cases, this implied that there was a dataset that contained mostly negative sentiment. In general, a bimodal distribution indicates that the

Dataset was not homogeneous and that underlying factors created two or more groups of data points with different characteristics. This could be useful information for further analysis and requires additional data exploration.

## 8 Future Works

We intend to conduct further qualitative data analysis to understand why some common ways to discern between real vs. false news did not hold in our Dataset. Since our datasets consisted primarily of news related to celebrities and politics, exploring this aspect could reveal insights into the relationship between truthfulness and news topics. Additionally, it could be interesting to explore merging names as individual items for the Apriori algorithm to find new rules.

Furthermore, we aim to investigate the cause of the bimodal distribution observed in the sentiment analysis results. By doing so, we hope to better understand the underlying patterns and characteristics of the news datasets that gave rise to this phenomenon. This analysis could provide valuable insights into the nuances of sentiment expression in news reporting and potentially inform future research in this area.

## REFERENCES

[1] L. Chiou and C.E. Tucker, "Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement," SSRN Electronic Journal, 2018, doi: 10.2139/ssrn.3209929.

[2] M. Gentzkow, J. Shapiro, and D. Stone. "Media bias in the marketplace: Theory," Technical report, National Bureau of Economic Research, 2014.

[3] A. Gupta et al., "Faking sandy: characterizing and identifying fake images on Twitter during hurricane sandy," Proceedings of the 22nd International Conference on World Wide Web, 2013.

[4] D. Vishwakarma, and C. Jain, "Recent state-of-the-art of fake news detection: A Review," 2020 International Conference for Emerging Technology (INCET), 2020.

[5] M. Szczepański, M. Pawlicki, R. Kozik et al., "New explainability method for BERT-based model in fake news detection," Scientific Reports, vol. 11, no. 1, 2021, article no. 23705.

[6] C. Castillo, M. Mendoza, and B. Poblete. "Information credibility on Twitter," Proceedings of the 22nd International Conference on World Wide Web, 2011.

[7] S. Kwon, et al., "Prominent features of rumor propagation in online social media," IEEE 13th International Conference on Data Mining, p. 1103–1108. IEEE, 2013.

[8] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media." arXiv, Mar. 27, 2019. Accessed: Mar. 23, 2023. [Online]. Available: http://arxiv.org/abs/1809.01286

[9] W. Y. Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection." arXiv, May 01, 2017. Accessed: Mar. 23, 2023. [Online].

[10] M. Mahyoob, J. Al-Garaady, and M. Alrahaili, "Linguistic-Based Detection of Fake News in Social Media," International Journal of English Linguistics, vol. 11, no. 1, DOI: 10.5539/ijel.v11n1p99, 2021.

[11] H. Reddy, N. Raj, M. Gala and A. Basava. Text-mining-based Fake News Detection Using Ensemble Methods. International Journal of Automation and Computing, vol. 17, no. 2, pp. 210-221, 2020. https://doi.org/10.1007/s11633-019-1216-5