DS 5500 – Final Report (Half-Semester Project)

Title: Is it all fake? – Fake News Detection

Authors: Gopalika Sharma and Surya Menon

GitHub: https://github.com/gsharma14/DS-5500-Fake-News

1 Summary

Recent events have truly highlighted the importance of consistent and accurate news reporting for our society to function effectively. With the prominence of the Internet, we are often bombarded with huge amounts of content that shape our world views and ultimately influence our behaviors. Access to high quality information is an essential resource, not only for individuals to make informed decisions in their daily lives, but also for communities to assess the actions of the leaders and institutions responsible for public safety and security.

While the significance and impact of news is constant, our news consumption habits have continued to evolve, especially within the past decade, due to the rapid rise of social media. These platforms allow for large amounts of information to be quickly accessed and easily shared to wide audiences. Consequently, recent studies have shown that close to 20% of American adults primarily receive political news from social media (1).

However, this democratization of information has also facilitated the growing influence of fake news in our mainstream culture (2). More content creates increasing opportunities for individuals to encounter and spread inaccurate stories to friends and family. As a result of our reliance on social media versus actual journalism, we have become less informed as communities and are often consuming content to validate our pre-conceived notions, rather than to educate. The large-scale negative impacts of fake news are still emerging, but some of the far-reaching consequences include: the dissemination of propaganda to wider audiences, attempts to undermine democratic systems, and the general erosion of trust in the government and media institutions (3).

The goal of this capstone project was to explore the features and text of news articles with machine learning and natural language processing (NLP) methods, in order to improve how individuals engage with this content and to reduce the impact of misleading stories. To achieve these goals, we used two datasets that contained news articles labeled as either “real” or “fake”. We first utilized the Fake News Inference Dataset (FNID), which is composed of political news stories between 2007 and August 2020 (4). We also created an article dataset covering the following topics: politics, entertainment, COVID-19, and disasters.

Extensive exploratory data analysis (EDA) was conducted on both datasets to identify general strategies that individuals can employ in their daily lives to evaluate article authenticity. We explored associated article metadata (e.g., publication source) and additional features, such article length, for differences in real and fake news content. We delved further into the raw text data through n-gram extraction, sentiment analysis, and topic modeling. Binary classification models for fake news detection were then developed as resources to effectively combat the onslaught of fake news stories in the widening news landscape. We also experimented with topic-specific classifiers and evaluated the performance between these models to assess if certain topics have more discernable patterns between real and fake content.

Our exploratory analysis revealed distinct patterns between fake and real news stories in terms of article length, publication source, punctuation, and sentiment. Individuals can leverage these findings to more effectively identify fake or misleading stories when browsing news content. Through predictive modeling, we were able to develop effective classifiers to detect general fake news stories; furthermore, topic-specific modeling revealed particular topics, such as politics and COVID-19 reporting, had higher accuracy in fake news detection, perhaps due to the clear differences in language patterns between the real and fake content for these topics.

2 Methods

2.1 Datasets and Preprocessing

To conduct EDA and general fake news classification we first worked with the FNID, from the IEEE DataPort platform, which contained over 17,000 news articles regarding United States (US) politics from a variety of

sources, including The New York Times, Twitter, and Fox News. These articles were web scraped by PolitiFact, a non-partisan organization that conducts fact-checking on primarily political news, and subsequently assigned labels of either "real" or "fake" (5). Additionally, this dataset included fields for the publication date, source, and the primary individual or organization on which the article was reporting.

We also created a combined topics dataset from a variety of sources with labeled articles in order to perform additional EDA and topic-specific fake news detection. Data was first pulled from FakeNewsNet, a repository developed at Arizona State University (ASU) that includes a set of political and gossip articles (6). Due to privacy policies, instead of a direct download, we recreated the full dataset utilizing the Twitter API and a Python script provided by ASU. Additionally, separate labeled datasets for both COVID-19 and disaster (i.e., natural disaster, accidents) reporting were found (7,8). Combining these sources resulted in a dataset, further described in Table A3 of the Appendix, with close to 30,000 articles. We observed that the majority of the data was entertainment news and that some class imbalance (i.e., count of real/fake articles) existed within topics.

To support NLP tasks and feature exploration, some general text preprocessing was performed. Specifically, we removed punctuation and non-alphanumeric characters, converted to lower-case, ran tokenization, and removed stop words (e.g., "a" or "the") by utilizing the Genism package in Python (9). This data cleaning was applied to the raw article text in both datasets, as well as to the *title* field in the combined topics data. Furthermore, for the *source* field in both datasets, we parsed the URL sources to extract the associated site domains.

2.2 Exploratory Data Analysis

2.2.1 Feature Analysis

To identify distinct patterns between real and fake articles and highlight potential contributing factors towards article authenticity, we looked at various article features (e.g., *source* and *date*) and extracted additional fields from the article text, such as article and title length. Histograms and bar plots were then generated to explore these features by article count frequency, as well as between fake and real text (10). We also examined the *date* and *speaker* fields exclusive to the FNID; the time feature in this dataset enabled us to investigate the trends in these article features over time. For the combined topics data, we performed an analysis on the *title* field, and also explored each news topic separately to see if particular subjects had different feature patterns.

Additionally, we were curious about differences in punctuation between real versus fake articles. Prior to initial text preprocessing, we looked at the counts of quotation marks and exclamation points present in both the real and fake text. We hypothesized that real articles would have higher counts of quotation marks, indicating the use of supporting information to back up the article content. We also theorized that fake news would contain more exclamations, which may indicate emotional content rather than factual reporting.

2.2.2 Text Analysis

Beyond feature analysis, raw text analysis was conducted on both datasets as well as on each news category in the combined topics data. Word clouds and n-gram analysis (i.e., frequencies of co-occurring words) were performed to find language patterns and content differences between real and fake article text (11).

Next, sentiment analysis was performed to identify the emotional tones of subjective information within news stories. Since fake news articles are often designed to push certain agendas through emotional manipulation, we believed that this content would contain more polarizing language. We first utilized the Valence Aware Dictionary for Sentiment Reasoning (VADER) in the NLTK library, which produced overall positive and negative values based on the sentiment scores of the words within the analyzed text; this model was created from a generalized sentiment lexicon and often works well on social media text (12). We also used the TextBlob library to quantify the magnitude of polarizing language used in an article and the text2emotion package to extract more specific emotions (e.g., anger, fear, sadness) within the articles (13,14).

Finally, topic modeling was conducted on the FNID to identify possible subtopics within the political news stories. Specifically, Latent Dirichlet allocation (LDA) was implemented and tuned to extract the appropriate number of topics in the data (15). These results helped identify other possible patterns in the text corpus that might affect our later modeling.

2.3 Modeling

2.3.1 General Fake News Detection

We initially focused on training classifiers to perform general fake news detection on both full datasets. To represent the text data in numeric form, document-term matrices were created with the TF-IDF (term frequency-inverse document frequency) scheme; these matrices represented the importance of each term in an article by combining the term frequency (TF) with the rareness of the term in the full corpus, or the inverse-document frequency (IDF). The number of words used as features was tuned (based on a maximum term frequency value) during model training in order to reduce the final input matrix size, while still providing enough information for classification (16). The datasets were then split into 80% training data and 20% test data.

Next, a series of general classifiers were fit on both datasets including Logistic Regression, Naïve Bayes classifier, Passive-Aggressive classifier (PA), and the following ensemble methods: Random Forest (RF), XGBoost ensemble (XGB), and gradient boosting (17). To optimize the various hyperparameters and improve performance, we implemented randomized grid search with cross-validation. The test accuracy, precision, and recall values were calculated to evaluate the final models.

In addition to these initial models, we also worked with the Bidirectional Encoder Representations from Transformers (BERT) and the Robustly optimized BERT (RoBERTa), which are pre-trained NLP models that have been optimized for language-based tasks. BERT was developed by Google in 2018 through bidirectionally training on a massive unlabeled text corpus in order to acquire a deeper understanding of linguistic patterns (18). Generally speaking, BERT masks a percentage of words in a sequence and then attempts to predict these masked words based on the context provided from surrounding words. This pre-training process allows for fine-tuning with subsequent training on additional input text (19). RoBERTa, an extension of BERT developed by Facebook, has been pre-trained on a larger amount of data over a longer period of time and introduces dynamic masking, which allows for variation in masking patterns as a technique to further enhance model training (20). We employed these models to achieve a possible boost in classification performance.

2.3.2 Topic-specific Fake News Detection

We also explored topic-specific fake news detection by splitting the combined topics data into the individual news categories—politics, gossip, COVID-19, and disasters—and training classifiers on each separate dataset. Due to the smaller datasets for each individual topic, we focused on general classifiers rather than the BERT models, which require more data to effectively train. There was some class label asymmetry for individual news topics, as seen in Figure A13 of the Appendix, so under-sampling was performed to balance each dataset. Similar to general fake news detection, Logistic Regression, Naïve Bayes, PA, RF, XGB, and gradient boosting models were trained on each topic and tuned utilizing cross-validation. The test accuracy, precision, and recall metrics were again reported, and performance was compared within and between topics. We were specifically interested in whether particular topics had language patterns that improved the classification results.

3 Results

3.1 Exploratory Data Analysis

3.1.1 FNID

When tracking the frequency of fake and real articles between 2007 and 2020, as seen in Figure A3 of the Appendix, we found that the quantity of fake news content continued to increase over time, particularly from 2016 onwards. We also observed that the rise in fake content roughly aligns with the steady increase of social media usage in the past decade (21). This trend also coincides with the highly polarizing 2016 election, during which social media platforms, particularly Facebook, were criticized for their role in spreading fake news.

Our analysis of article length revealed that on average both fake and real text were similar in length, with most articles being between 500 and 1000 words. However, the distribution of word counts in real and fake articles in Figure A5 of the Appendix shows that fake stories tended to be skewed further right, which indicates a slightly higher tendency for fake content to be longer than real news.

We also found that the speaker or organization that an article is about often indicated whether the content was real or misleading. Figure A1 of the Appendix displays the frequency of real and fake articles associated with the the past two US presidents and the current president-elect: Barack Obama, Donald Trump, and Joe Biden. We found that Trump was mentioned significantly more in fake articles compared to his counterparts. Although this particular finding may indicate a political bias, exploring the full list of speakers in Figure A4 of the Appendix shows that speakers in both political parties were highly represented in real articles.

The publication source analysis results in Figure 1 indicate that the article source may be integral in determining the authenticity of a news report. Many social media sites, specifically Facebook and Twitter, were the top purveyors of fake news stories whereas traditional news sources (e.g., ABC News, The Washington Post) were producing most of the real articles. Interestingly, we also noticed that YouTube was a top provider of both real and fake content, which is likely due to the wide variety of video content (from official news clips to amateur videos) being uploaded to this platform on a regular basis. When investigating further, we discovered that many sources often created both real and fake news content. Figure A6 of the Appendix illustrates that even "trusted" news organizations, such as The New York Times or C-SPAN, were just as culpable of spreading fake content at some point. Ultimately, there is no "perfect" source.

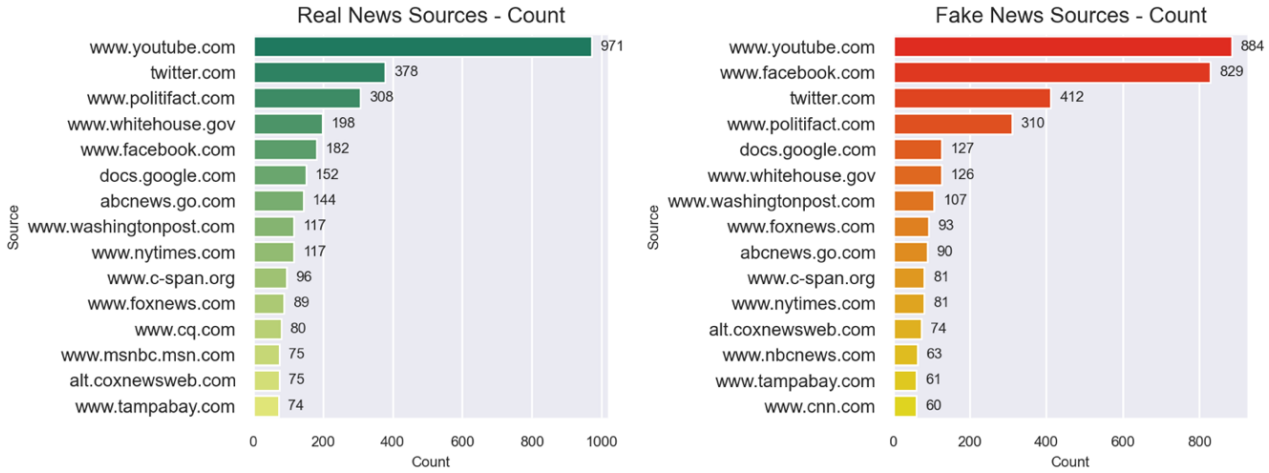


Figure 1: Top 15 Real and Fake News Sources

Our sentiment analysis results in Figure 2 show that fake articles generally had more overall negativity compared to real news stories. When exploring this trend over time in Figure 3, we found that fake stories were also more negative on average than real content for most of the dataset timespan, with larger differences in average negativity after 2016. Additionally, Figure A8 of the Appendix confirms that fake stories utilized more angry sentiments on average compared to real stories, with particular spikes in the last two years.

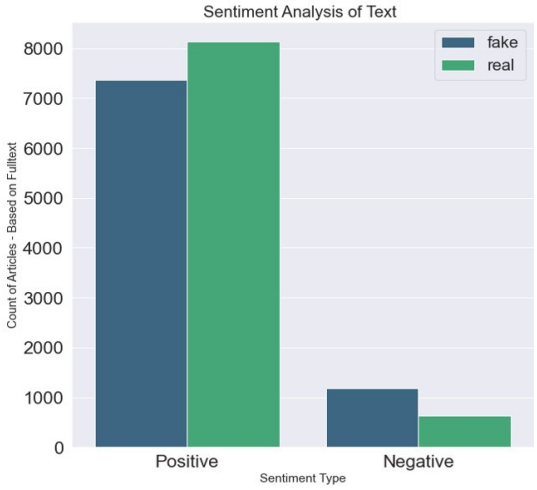


Figure 2: Overall Sentiment Article Count

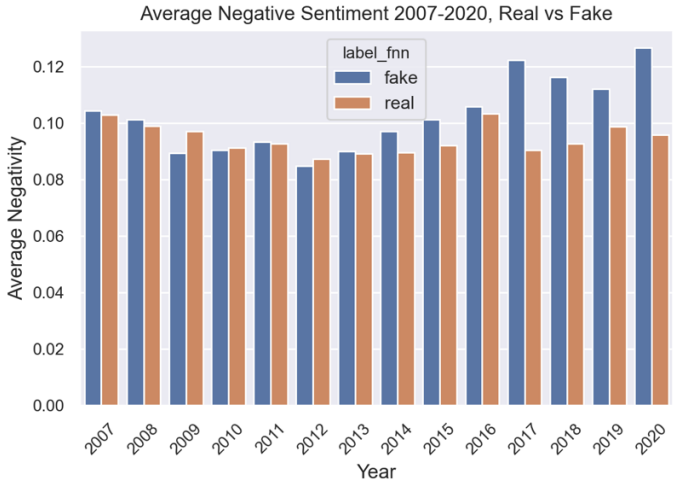


Figure 3: Average Negative Sentiment, 2007-2020

Our LDA results identified 15 relevant political subtopics including energy, immigration, healthcare, and elections. The language patterns within these categories may impact how well we can classify real and fake news stories in this dataset, as certain subtopics (e.g., elections) may have clearer distinctions between real and fake content compared to others.

3.1.2 Combined Topics Dataset

Figure 4 shows the distribution of article title length in the combined topics dataset and reveals a clear distinction between real and fake articles. Specifically, we observed that fake news stories frequently had article titles much shorter in length compared to those of real articles. Additionally, through raw text analysis we found that fake news titles employed more capitalization and exclamation points and made use of fewer stop words.

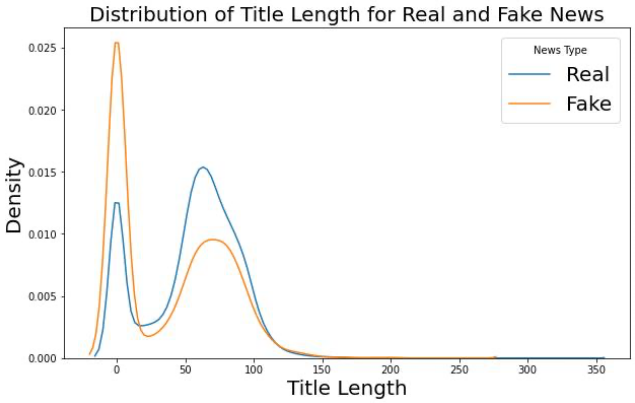


Figure 4: Article Title Length Distribution

When exploring the top bigrams in this dataset, which was mostly composed of entertainment news, we found significant overlap between fake and real article text; Figure A15 of the Appendix shows that many of the top words in both news types contained celebrity names and their locations. We also found that fake stories frequently mentioned ambiguous sources (e.g., "source told" or "source close"), often as a method to back up alleged claims about public figures that tended to have little concrete evidence.

Our analysis of quotations marks further explored the use of verified sources and references to support news reporting and confirmed that real stories generally had much higher instances of quotations compared to fake stories. Tables 1 and 2 show that for both article title and body text, real articles employed more quotations marks in total as well as on average.

News Type	Total Number of Quotations	Mean Number of Quotations
Real	110613	5.89
Fake	33014	3.20

Table 1: Article Text Quotation Statistics

News Type	Total Number of Quotations	Mean Number of Quotations
Real	763	0.041
Fake	236	0.023

Table 2: Article Title Quotation Statistics

From our brief analysis of individual topics in this dataset, we found that for political news, fake stories had longer article lengths and on average expressed more negativity compared to real content. These patterns strengthened the similar observations identified in the FNID. When exploring gossip stories, we found significant overlap between real and fake news publications; sources producing "real" entertainment news (e.g., People) were often also distributing many instances of fake content as well.

Finally, our analysis of COVID-19 reporting revealed clear distinctions between real and fake news reporting. Figure A20 of the Appendix shows the top sources for this topic by article count and found that most of the top real publications were academic and scientific institutions (e.g., Center for Disease Control, Harvard University), while fake sources included primarily Facebook and many sites that appear political in nature. The top bigrams for this topic further exposed distinct patterns; Figure 5 demonstrates that while real stories were focused on imparting essential information regarding public safety (e.g., social distancing and hand washing), fake content utilized much more polarizing language, such as "biological weapon" and "warfare".

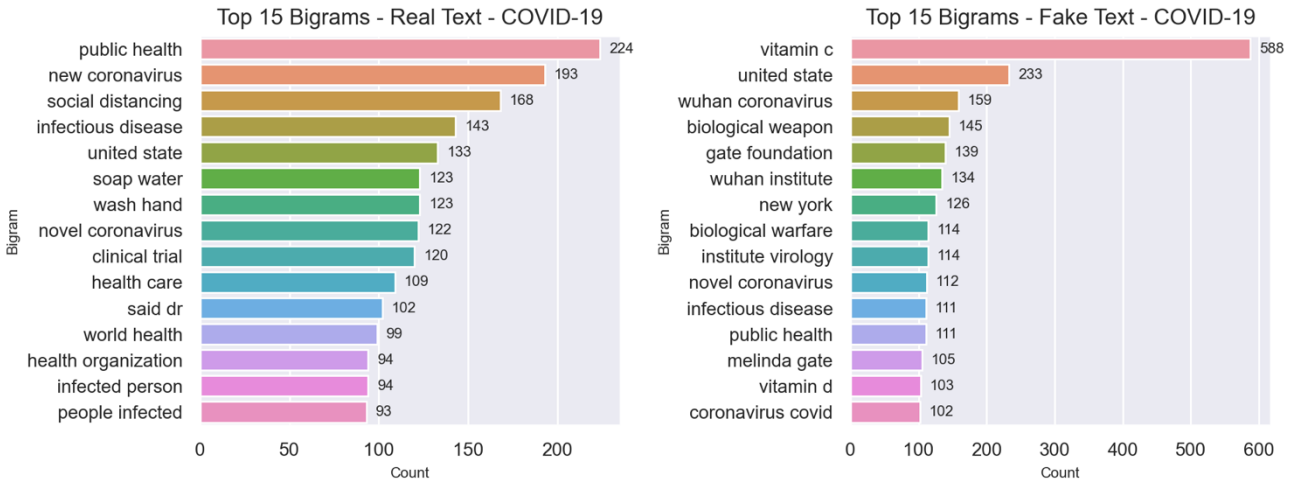


Figure 5: Top 15 Bigrams, COVID-19 News

3.2 Modeling

3.2.1 General Fake News Detection

When predicting fake news in the FNID, the optimal model after hyperparameter tuning was the XGB model, as seen in Table 3 and Figure 6. This model only achieved about 72% test accuracy, with similar precision and recall metrics. Perhaps the distinct subtopics within the data, as seen in our LDA results, impacted the ability to detect general language patterns between real and fake content. Additionally, neither the BERT nor RoBERTa models achieved improved performance compared to the general classifiers. Training on larger datasets and for additional time could help these models better detect language patterns and more effectively detect fake content. Furthermore, additional training on our text corpus in order to learn word embeddings specific to the dataset may also help boost the performance of these pre-trained models.

Classifier	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.72	0.72	0.715
Naïve Bayes	0.65	0.66	0.65
Passive Aggressive	0.68	0.67	0.67
Random Forest	0.70	0.71	0.695
Gradient Boosting	0.71	0.72	0.71
XGBoost	0.72	0.72	0.72
BERT	0.64	0.65	0.63
RoBERTa	0.69	0.69	0.68

Table 3: FNID Fake News Detection Metrics

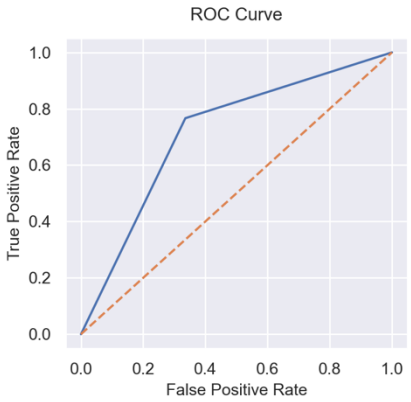


Figure 6: FNID – XGB ROC Curve

With fake news detection on the full combined topics data, we observed significant improvement compared to the FNID when trained on the same classifiers. Table A6 of the Appendix reveals that Logistic Regression had the best performance, with about 80% test accuracy. These results indicate that some news topics may be easier to extract patterns from compared to others; similar to our EDA results, differences between fake and real news appeared more subtle in the FNID compared to the combined topics data. Additionally, this dataset incorporated articles titles into the model input, which may have also contributed towards the performance results.

3.2.2 Topic-specific Fake News Detection

Table 4 shows that detecting fake political articles had improved classification compared to the full topics data, with XGB having the best performance. We also found that detecting fake COVID-19 articles resulted in the best classification results across all datasets, with 93% test accuracy for the PA model, as seen in Table 5. While

the COVID-19 dataset was limited to articles from January to August 2020, our EDA revealed clear differences in the language patterns for real and fake content, which may have contributed towards the classifiers achieving optimal performance. Incorporating additional data over time may help further refine these results.

Classifier	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.87	0.87	0.87
Naïve Bayes	0.76	0.79	0.77
Passive Aggressive	0.88	0.88	0.88
Random Forest	0.83	0.84	0.83
Gradient Boosting	0.81	0.82	0.81
XGBoost	0.89	0.88	0.89

Table 4: Politics Fake News Detection Metrics

Classifier	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.92	0.92	0.915
Naïve Bayes	0.89	0.89	0.89
Passive Aggressive	0.93	0.94	0.93
Random Forest	0.90	0.91	0.90
Gradient Boosting	0.86	0.87	0.86
XGBoost	0.86	0.86	0.86

Table 5: COVID-19 Fake News Detection Metrics

Our gossip fake news detection, as seen in Table A7 of the Appendix, produced similar performance results to the full dataset across most models and evaluation metrics. Perhaps this topic had lower relative performance due to the large overlap in content and language between the real and fake stories in this domain, which made it harder to identify distinct language patterns.

Finally, our disasters fake news detection had the lowest performance compared to other individual topics. Table A8 of the Appendix reveals that Logistic Regression had about 81% test accuracy, which showed slightly better performance compared to other classifiers. We hypothesized that the ability to detect clear patterns between fake and real content may be affected by the wide scope of information and subtopics (from natural disasters to car accidents) within the disasters data, which is similar to our FNID modeling findings.

4 Discussion

4.1 Conclusions

This capstone project intended to improve how individuals evaluate the authenticity of the news content they consume. Through exploratory data analysis, we sought to develop strategies that individuals can use to better filter out misleading content. Some of the key approaches include:

- *Question the source*: The credibility of a publication is often essential in determining whether an article contains trustworthy information, but even reliable news sources can output misleading content.
- *Assess your emotional response*: Fake news is often designed to manipulate emotions through the use of buzz words and polarizing language to evoke fear and negative feelings.
- *Check the headlines*: Article titles associated with fake content tend to be shorter and utilize evocative words to lure an individual into reading a story without facts to support claims made in the headline.
- *Track punctuation*: Quotations marks may be a good indicator of article authenticity, as they can signal the existence of supporting information. In contrast, exclamation points may be a sign of emotional content and opinions rather than facts, particularly for politics and public health news.

Another general tip we encountered from our research is to check the web domain of an article to determine if the news source is attempting to impersonate a credible publication (e.g., abcnews.com.co trying to imitate ABC News). Additionally, finding multiple articles and sources on a topic may also be helpful to fully understand a topic and establish a comprehensive representation of the truth.

In our attempts to detect fake news stories utilizing machine learning and NLP methods, we were able to develop effective classifiers, particularly for identifying misleading political and health-related stories. These topics generally had clear differences in language patterns and credible sources, which facilitated the classifiers in distinguishing between real and fake content.

4.2 Timeline

We were able to successfully complete our project goals and milestones on time. However, due to the addition of the combined topics dataset, which necessitated additional time for conducting EDA, as well as the training time required for the RoBERTa models, we were not able to achieve our stretch goal of creating a web application to classify user-submitted text input as real or fake.

4.3 Future Work

In addition to implementing a web application, we hope to further improve our project by refining our BERT and RoBERTa classifiers with additional training and text data in the future. We also believe that incorporating features extracted from our EDA, such as article length, as additional inputs may improve prediction performance by capturing other text pattern variations between real and fake news content. We would also be interested in seeing how multiple sources report on the same news story to establish credibility ratings for different publications; specifically, if news sources circulate the same fake content, we can more easily identify and expose questionable sources.

One interesting extension to this project would be to focus more on social media posts and the fake news patterns specific to these platforms; additionally, identifying fake and "bot" accounts may be another important avenue to reduce the large outflow of fake news content.

Furthermore, exploring political bias in relation to credible news sources may be critical in improving public trust in news reporting. It is important to acknowledge that misleading content does not stem from one political party or ideology and to then identify accurate sources of information from multiple political perspectives.

Finally, with the increasing trend of "deepfake" media that leverage artificial intelligence (AI) methods to swap a person in an image or video with another person's likeness, exploring fake news beyond text is an important future step. This content is incredibly deceptive and can easily spread a lot of false information and confusion. Accurately identifying this technology remains an open challenge, as many current algorithms have low accuracy in detecting unseen cases (22).

Through our work on this project, we have determined that there is no single solution to identify and filter out fake news. Automatic detection through machine learning can be a useful tool but is not infallible, and algorithms must be re-trained to capture novel variations in fake news content over time. In addition to fake news detection resources, providing individuals with the skills to better evaluate news content may be essential in more quickly identifying misleading stories even as this content gets adapted into different forms. We hope that our work can help reestablish journalistic standards for consistent and trustworthy news reporting and ultimately reduce the impact of the fake news content that has steadily seeped into our society.

5 Statement of contributions

Both team members contributed to all aspects of the project (e.g., EDA and pre-processing, modeling) due to the small group size. To complete this project, the team members made the following contributions:

1. *Gopalika Sharma* – conducted EDA on both the FNID and combined topics dataset, refined and displayed topic modeling results for the FNID, performed predictive modeling on both datasets, and worked on fine-tuning the BERT and RoBERTa models.
2. *Surya Menon* – created the combined topics data from multiple data sources and using the Twitter API, conducted EDA on both the FNID and combined topics dataset including an analysis of changes in the FNID patterns over time, performed predictive modeling on both datasets, and tuned the topic-specific classifier models.

References

1. Mitchell, A., Jurkowitz, M., Oliphant, B. J., & Shearer, E. (2020, July 30). Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable. Pew Research Center. https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/
2. Flores, L. (2020, September 30). Fake News: Why Does it Persist and Who's Sharing it? The Decision Lab. https://thedecisionlab.com/insights/society/fake-news-why-does-it-persist-and-whos-sharing-it/
3. West, D. M. (2017, December 18). How to combat fake news and disinformation. Brookings. https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/
4. Sadeghi, F., Bidgoly, A. J., & Amirkhani, H. (2020, August 18). FNID: Fake News Inference Dataset. IEEE. https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset
5. The Poynter Institute. (n.d.). PolitiFact. PolitiFact. Retrieved 2020, from https://www.politifact.com/
6. KaiDMML/FakeNewsNet. (n.d.). GitHub. Retrieved 2020, from https://github.com/KaiDMML/FakeNewsNet
7. susanli2016/NLP-with-Python. (n.d.). GitHub. Retrieved 2020, from https://github.com/susanli2016/NLP-with-Python/tree/master/data
8. Kaggle. (n.d.). Real or Not? NLP with Disaster Tweets | Kaggle. Retrieved 2020, from https://www.kaggle.com/c/nlp-getting-started
9. Gensim. (n.d.). Gensim: topic modelling for humans. Retrieved 2020, from https://radimrehurek.com/gensim/auto_examples/index.html#documentation
10. BuzzFeed News Analysis and Classification. (2020, October 20). Kaggle. https://www.kaggle.com/sohamohajeri/buzzfeed-news-analysis-and-classification
11. Gandhi, R. (2020, May 8). Getting Real with Fake News - Towards Data Science. Medium. https://towardsdatascience.com/getting-real-with-fake-news-d4bc033eb38a
12. Pandey, P. (2018, September 23). Simplifying Sentiment Analysis using VADER in Python (on Social Media Text). Medium. https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f
13. Tutorial: Quickstart — TextBlob 0.16.0 documentation. (n.d.). TextBlob. Retrieved 2020, from https://textblob.readthedocs.io/en/dev/quickstart.html
14. Band, A. (2020, September 14). Text2emotion: Python package to detect emotions from textual data. Medium. https://towardsdatascience.com/text2emotion-python-package-to-detect-emotions-from-textual-data-b2e7b7ce1153
15. Fake News Detection on Twitter EDA. (2020, April 29). Kaggle. https://www.kaggle.com/hamditarek/fake-news-detection-on-twitter-eda
16. DataFlair. (2020, August 6). Advanced Python Project – Detecting Fake News with Python. DataFlair. https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/
17. Dounis, F. (2020, June 5). Detecting Fake News With Python And Machine Learning. Medium. https://medium.com/swlh/detecting-fake-news-with-python-and-machine-learning-f78421d29a06
18. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language. https://arxiv.org/pdf/1810.04805.pdf
19. Alammar, J. (2018, December 2). The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning). GitHub - Jay Alammar. http://jalammar.github.io/illustrated-bert/
20. Khan, S. (2020, September 4). BERT, RoBERTa, DistilBERT, XLNet — which one to use? Medium. https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8
21. Perrin, A., & Anderson, M. (2019, April 10). Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/
22. Knight, W. (2020, September 15). Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them. Wired. https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/

Appendix

Exploratory Data Analysis – FNID

Feature	Description	Type
Date	Article Date	Interval
Speaker	Person/organization to whom text relates	Nominal
Sources	Article publication	Nominal
Article Text	Article raw text	Nominal
Fake News Class (label_fnn)	Real or Fake	Nominal

Table A1: FNID Features

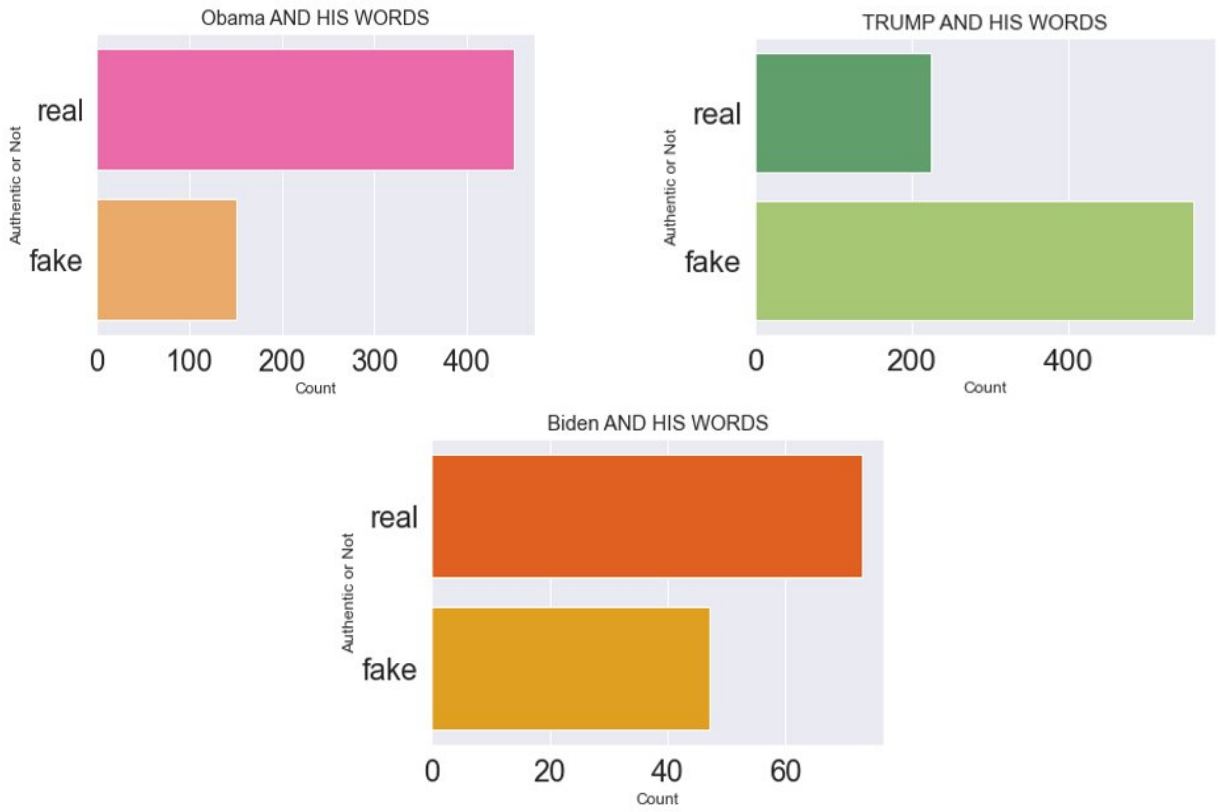


Figure A1: FNID – Article Count for Presidential Speakers, 2007-2020

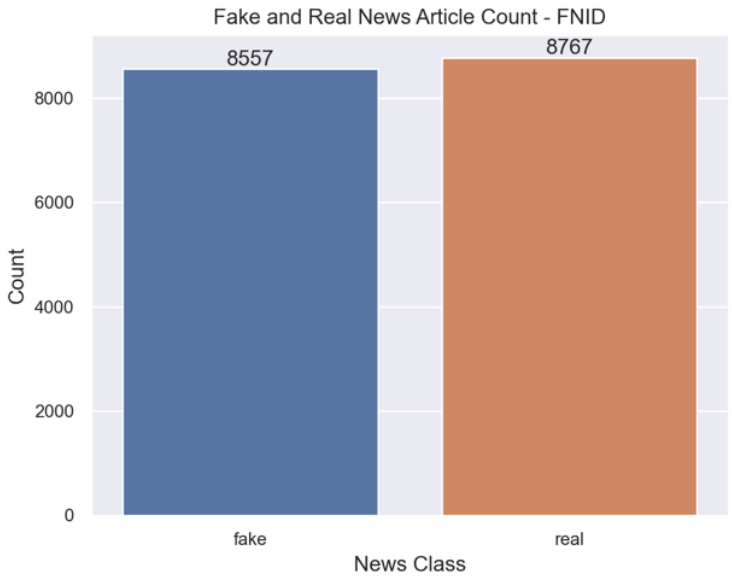


Figure A2: FNID – Fake and Real Article Count

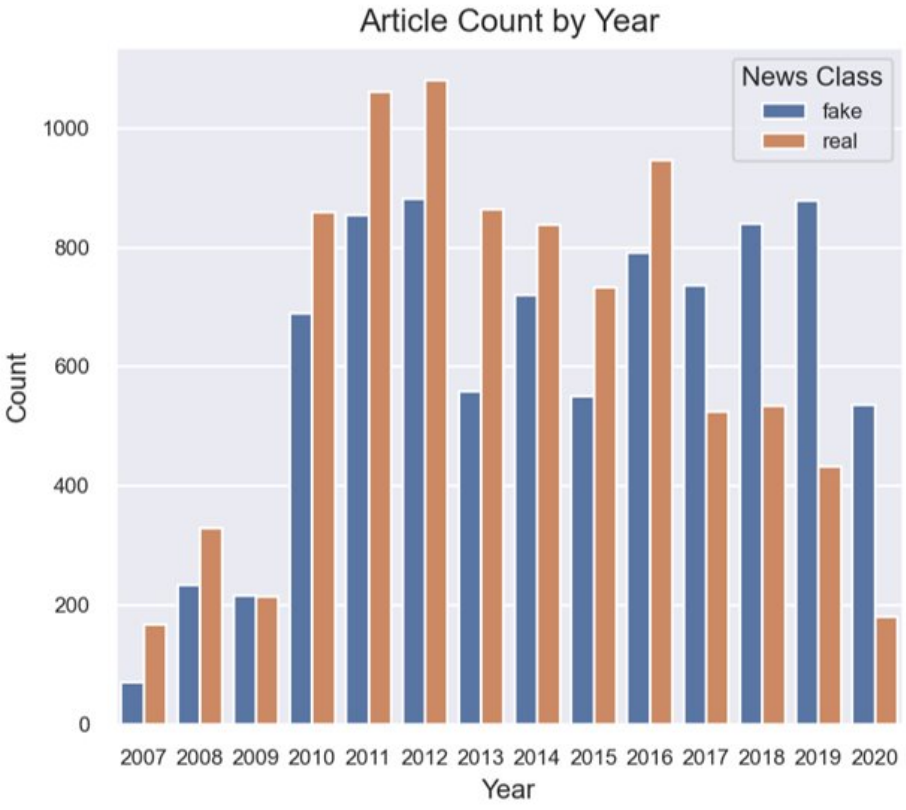


Figure A3: FNID – Article Count by Year, 2007-2020

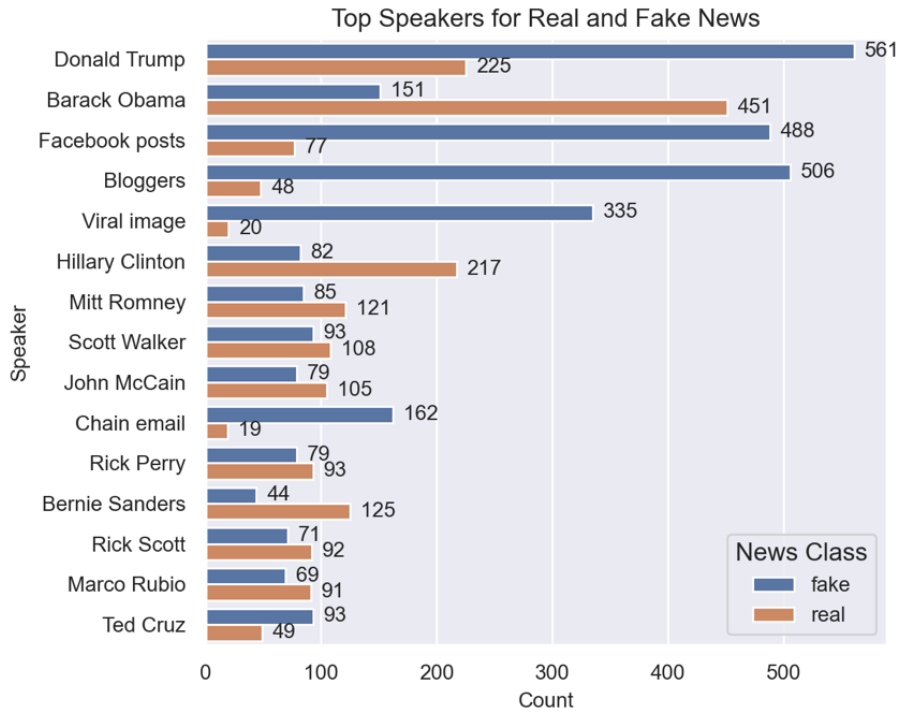


Figure A4: FNID – Top Speakers for Real and Fake News

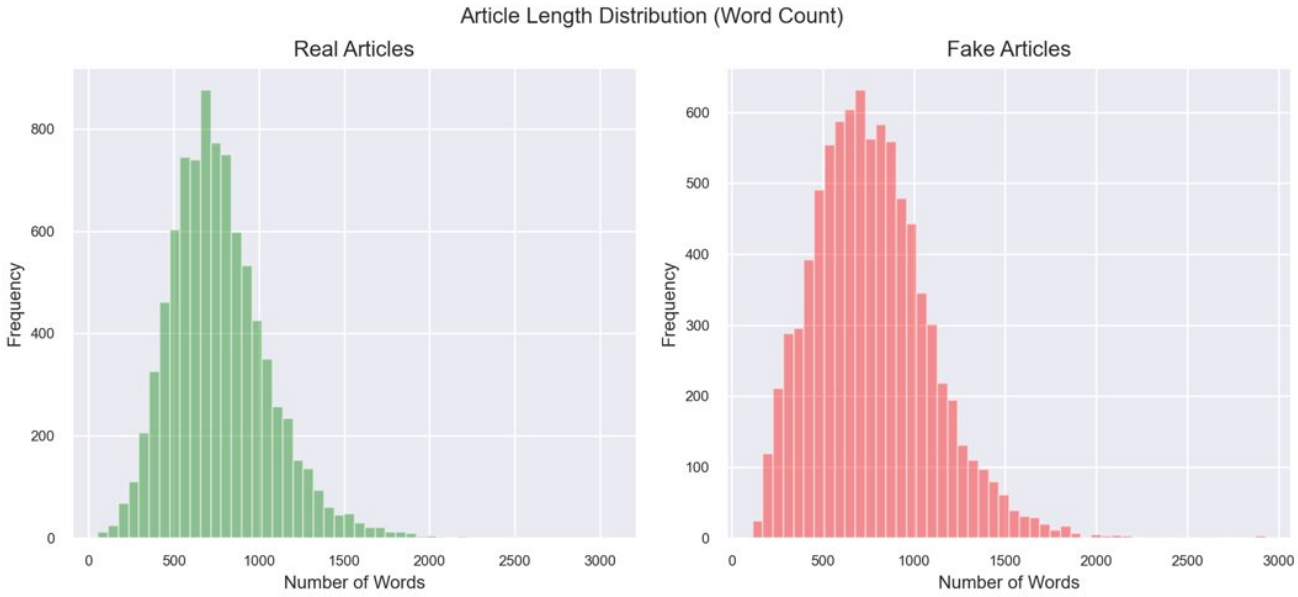


Figure A5: FNID – Article Length Distribution for Real and Fake Articles

News Type	Total Number of Quotations	Mean Number of Quotations	Total Number of Exclamations	Mean Number of Exclamations
Real	40850	4.66	652	0.074
Fake	40213	4.70	1794	0.21

Table A2: FNID – Article Text Punctuation Statistics

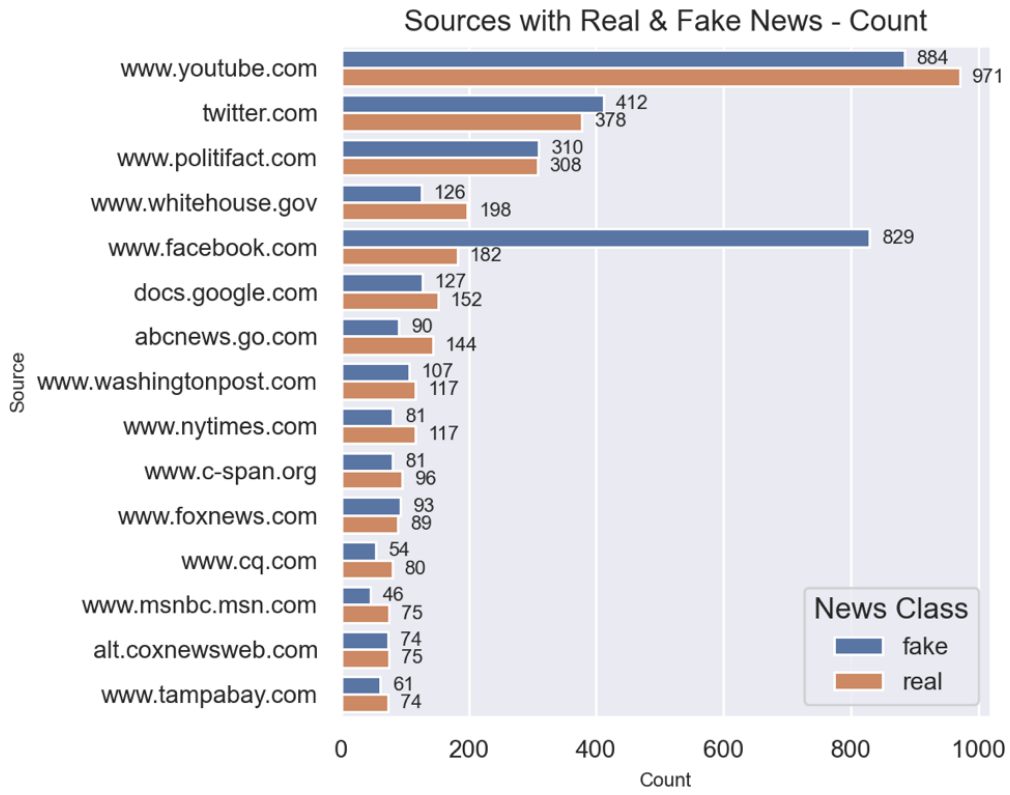


Figure A6: FNID – Sources with Real and Fake News – Article Count

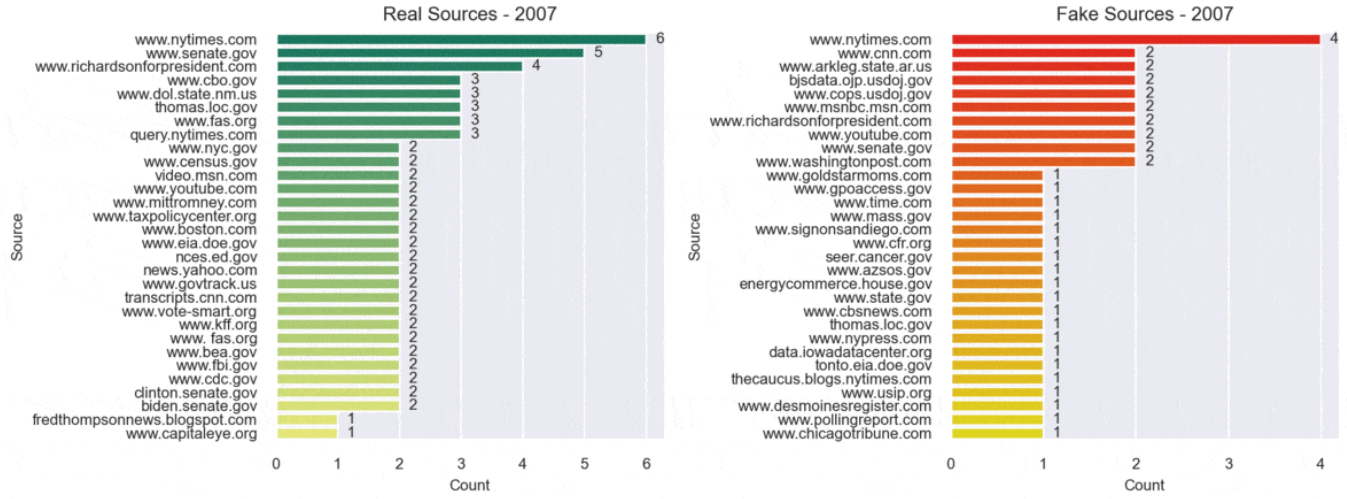


Figure A7: FNID – Change of Real and Fake Sources, 2007-2020*

*GIF file available in EDA-FNID folder of GitHub repository

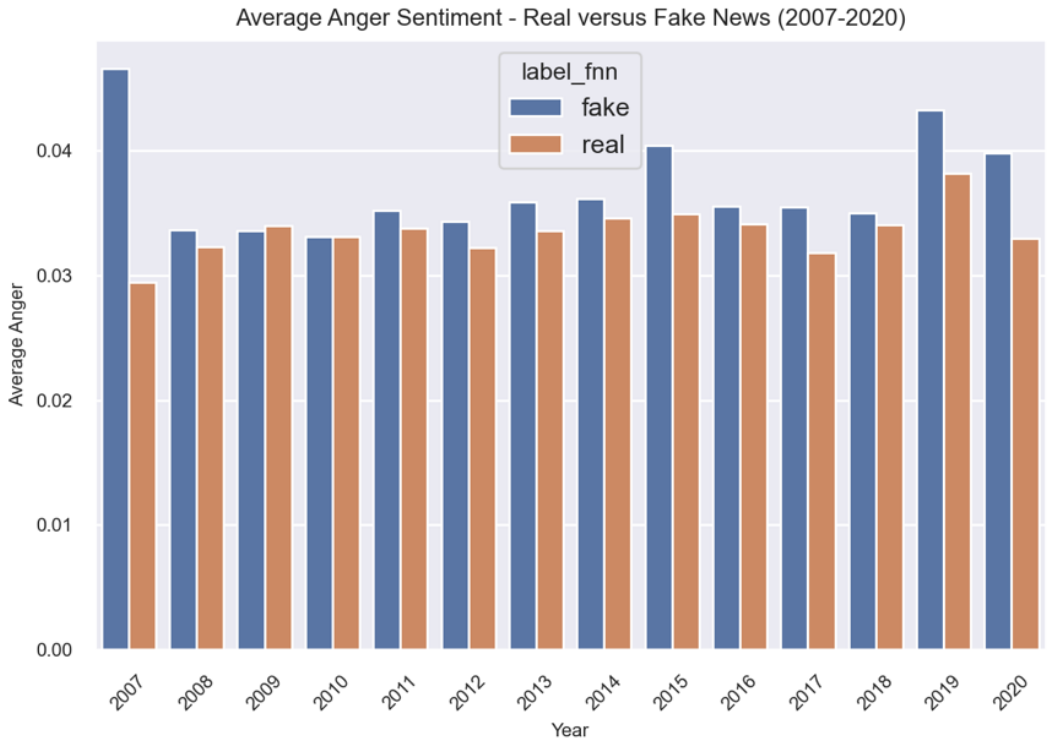


Figure A8: FNID – Average Anger Sentiment, 2007-2020

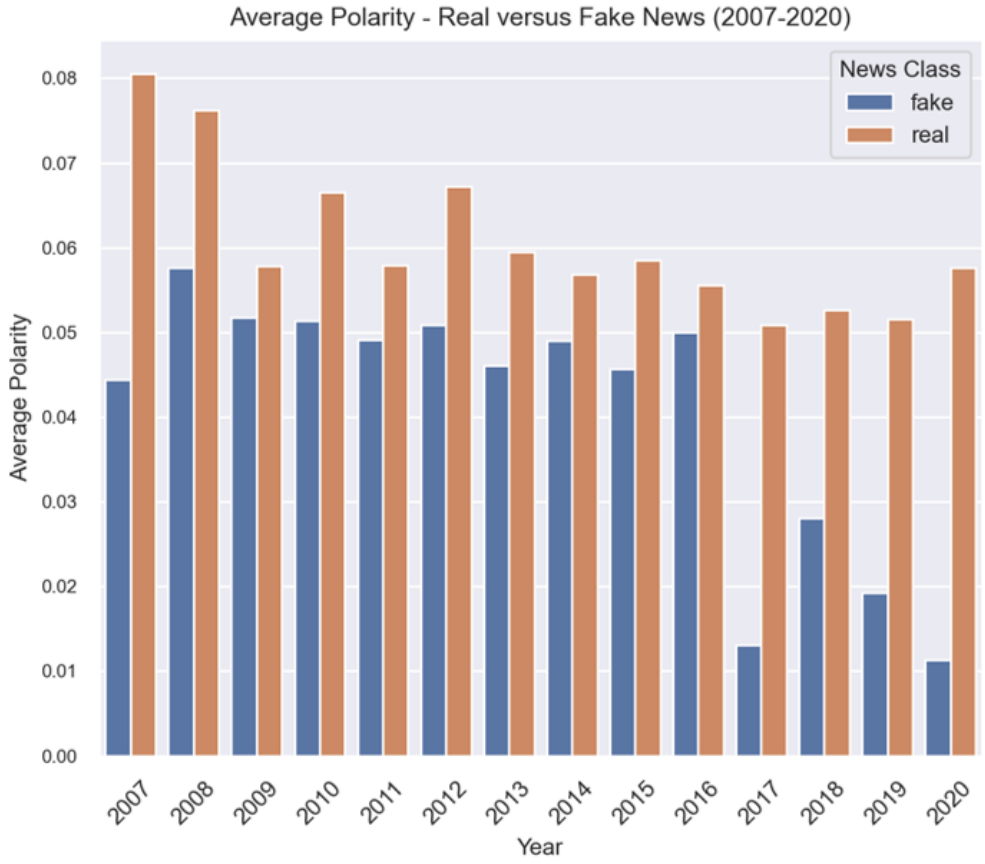


Figure A9: FNID – Average Article Polarity, 2007-2020

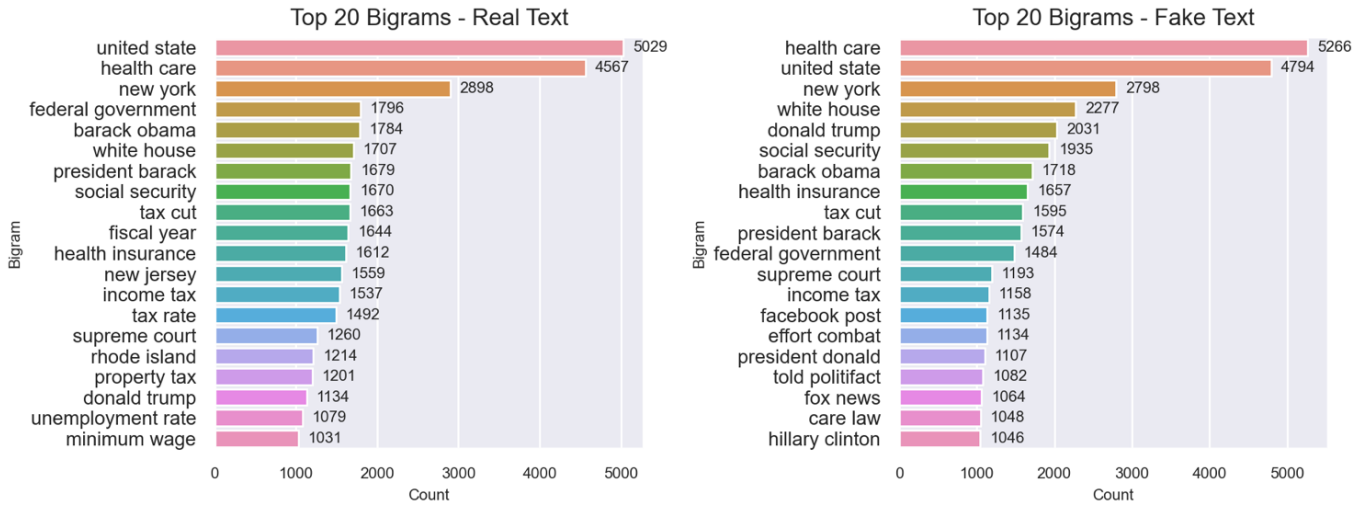


Figure A10: FNID – Top 20 Real and Fake Article Bigrams

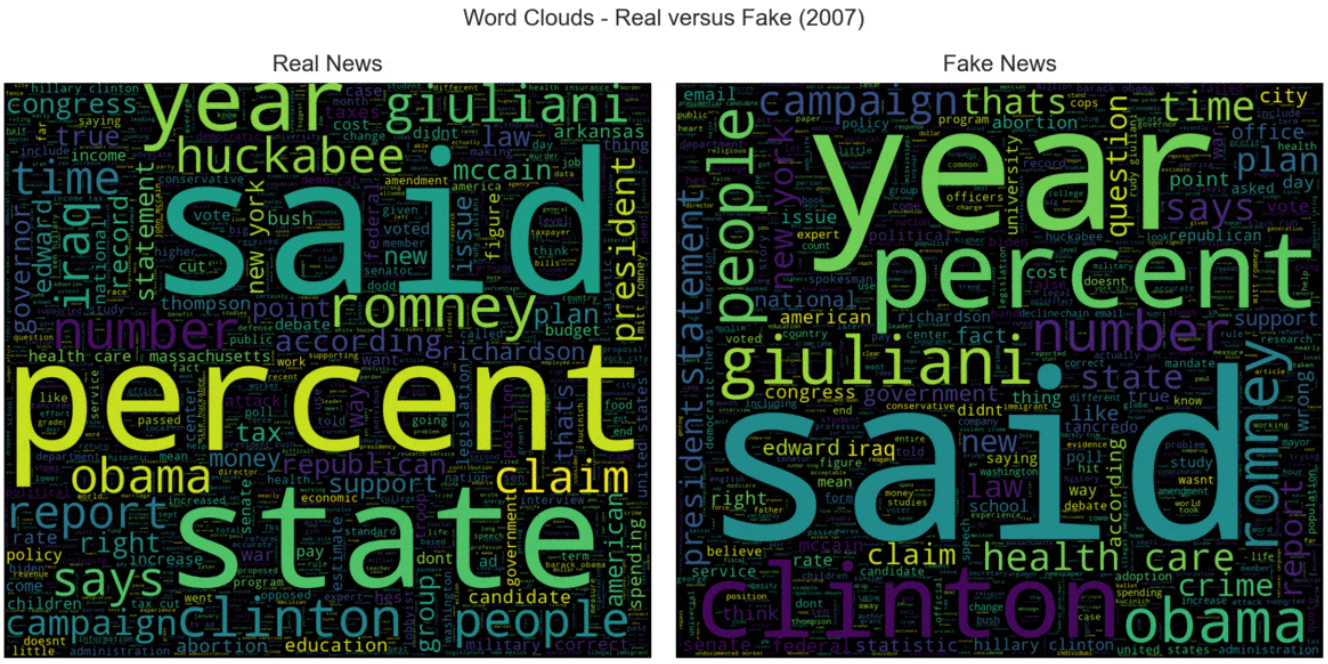


Figure A11: FNID – Word Clouds – Real and Fake Text, 2007-2020*

*GIF file available in EDA-FNID folder of GitHub repository

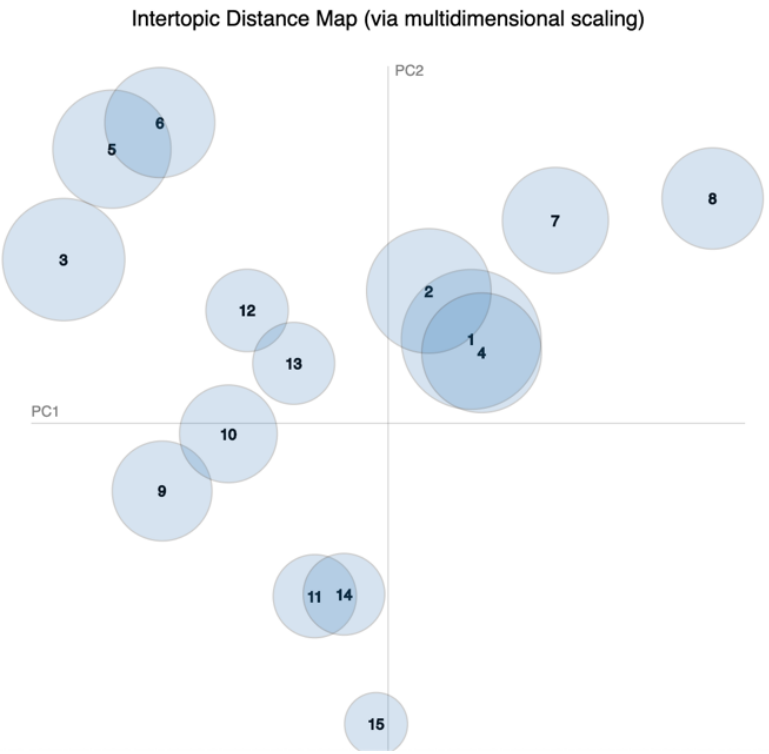


Figure A12: FNID – Topic Modeling subtopics*

*Interactive HTML file available in EDA-FNID folder of GitHub repository

Exploratory Data Analysis – Combined Topics Dataset

Feature	Description	Type
News Topic (ID)	5 values	Nominal
Title	Article title (if provided)	Nominal
Article Text	Article raw text	Nominal
Source	Article publication	Nominal
Fake News Class (target)	Real or Fake	Nominal

Table A3: Combined Topics Dataset Features

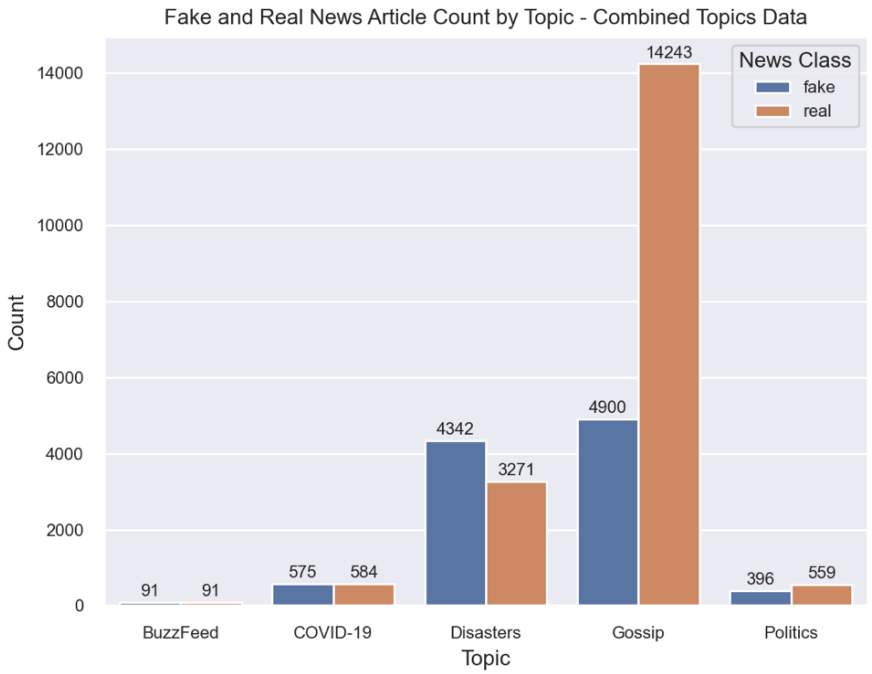


Figure A13: Combined Topics – Fake and Real Article Count

News Type	Total Number of Exclamations	Mean Number of Exclamations
Real	23198	0.951
Fake	9794	1.24

Table A4: Article Text Exclamation Statistics

News Type	Total Number of Exclamations	Mean Number of Exclamations
Real	768	0.041
Fake	437	0.042

Table A5: Article Title Exclamation Statistics

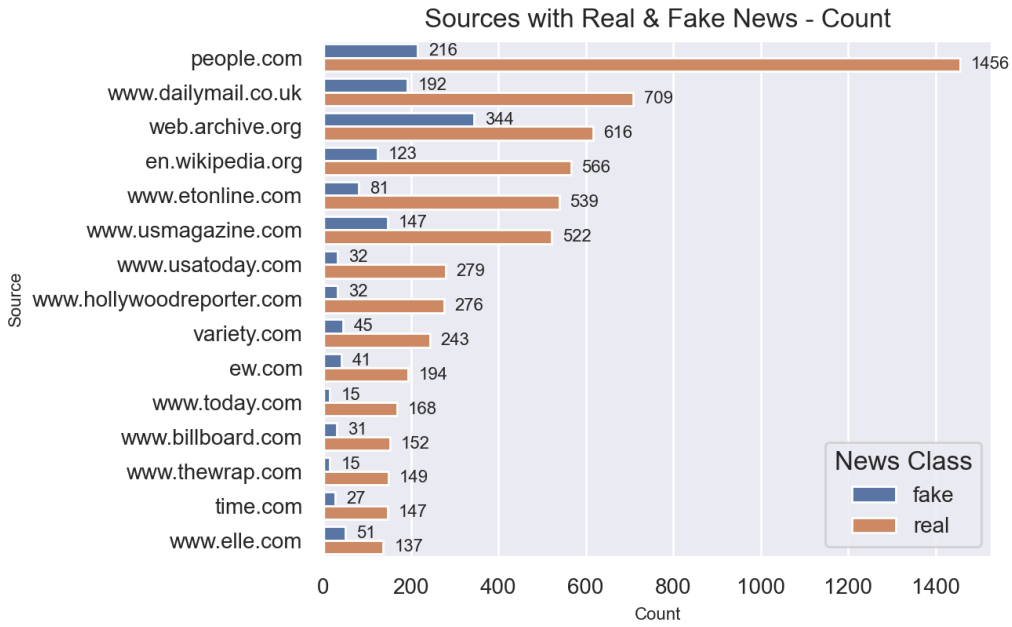


Figure A14: Combined Topics – Sources with Real and Fake News – Article Count

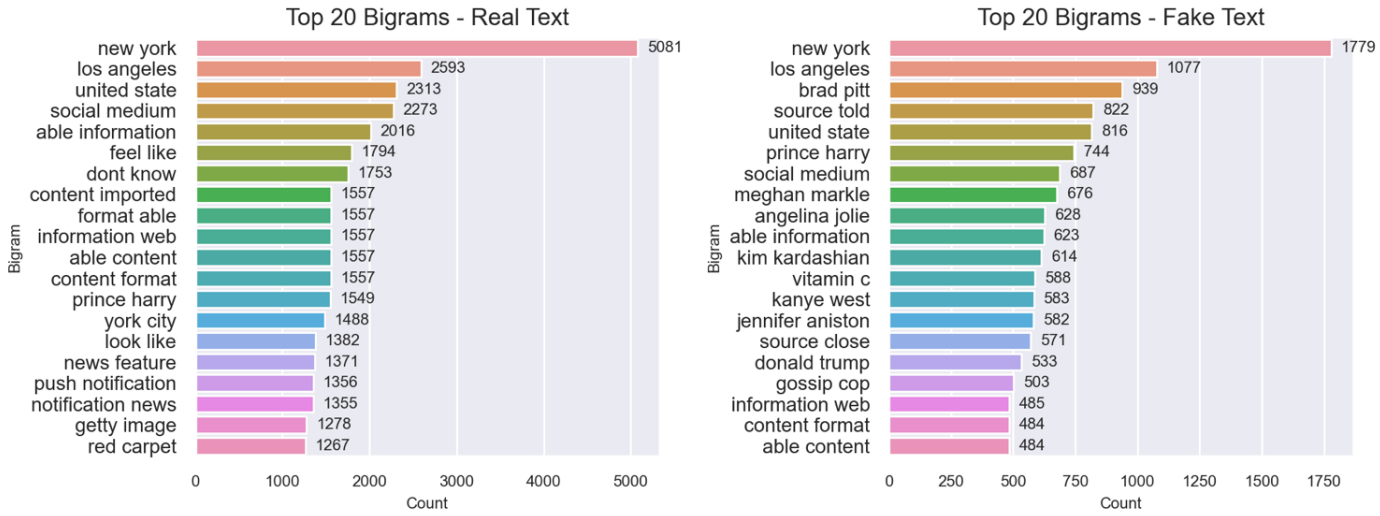


Figure A15: Combined Topics – Top 20 Real and Fake Article Bigrams

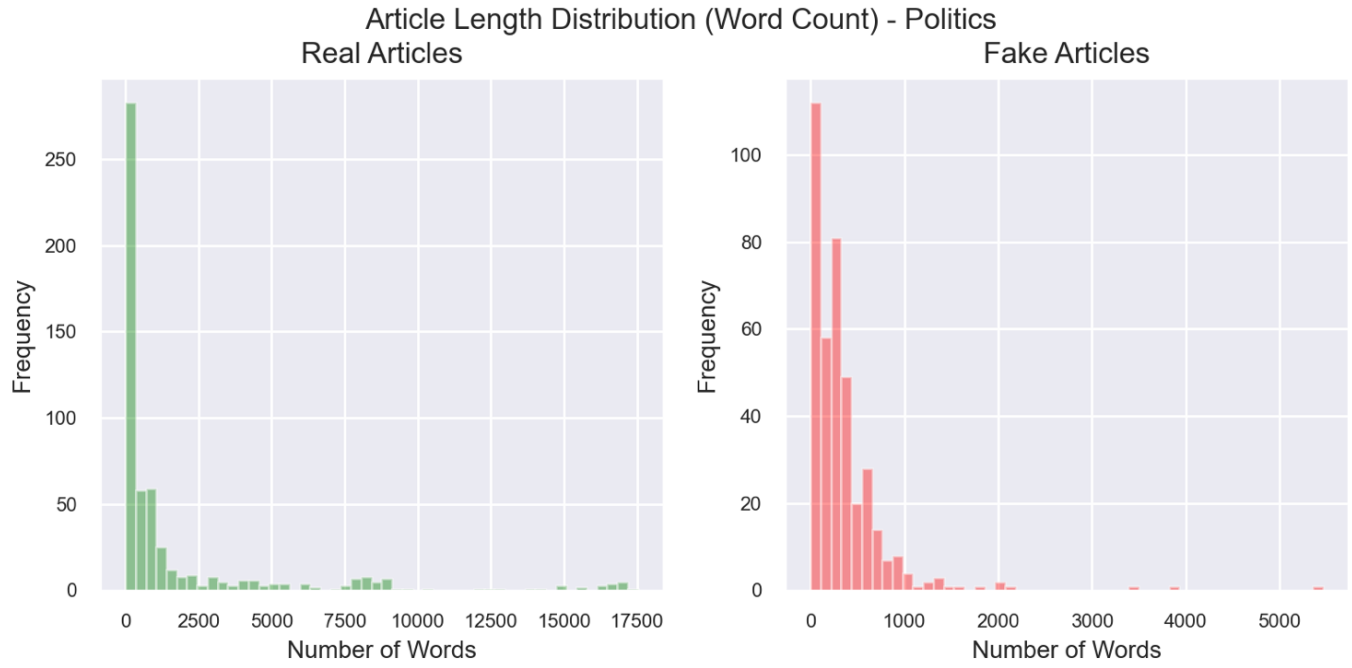
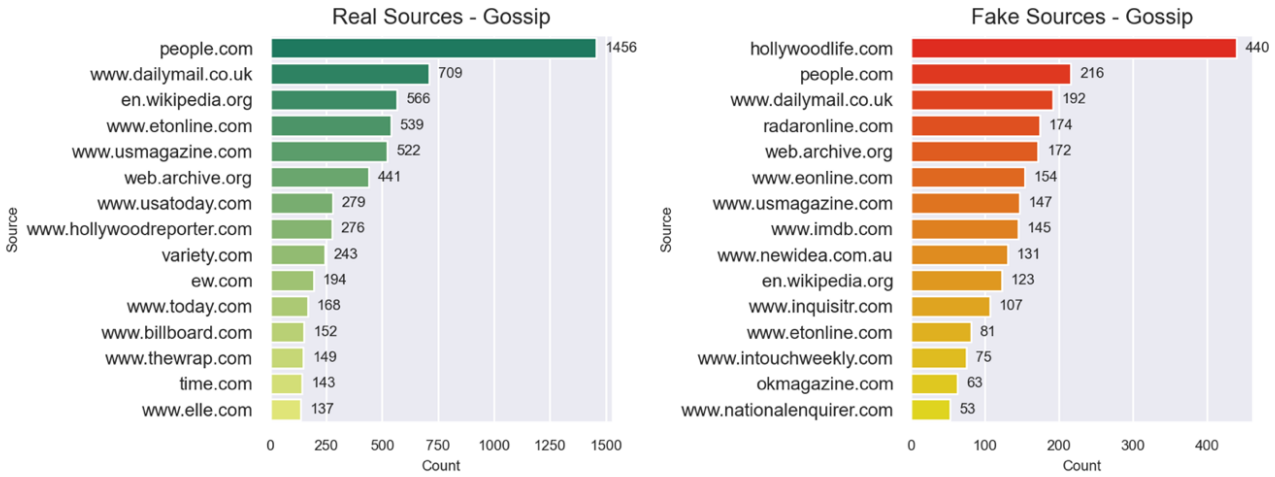


Figure A16: Political Article Length Distribution for Real and Fake Articles

Figure A17: Political Article Negativity Distribution for Real and Fake Articles



A18: Top 15 Real and Fake News Sources – Gossip Articles

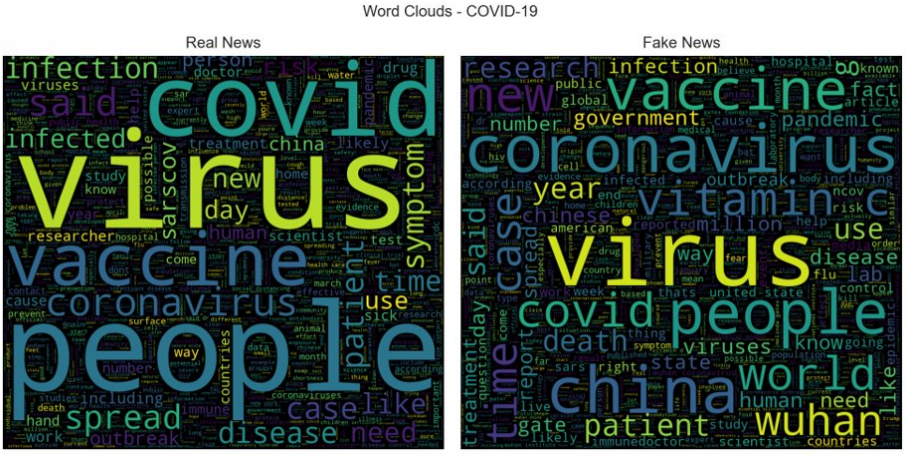
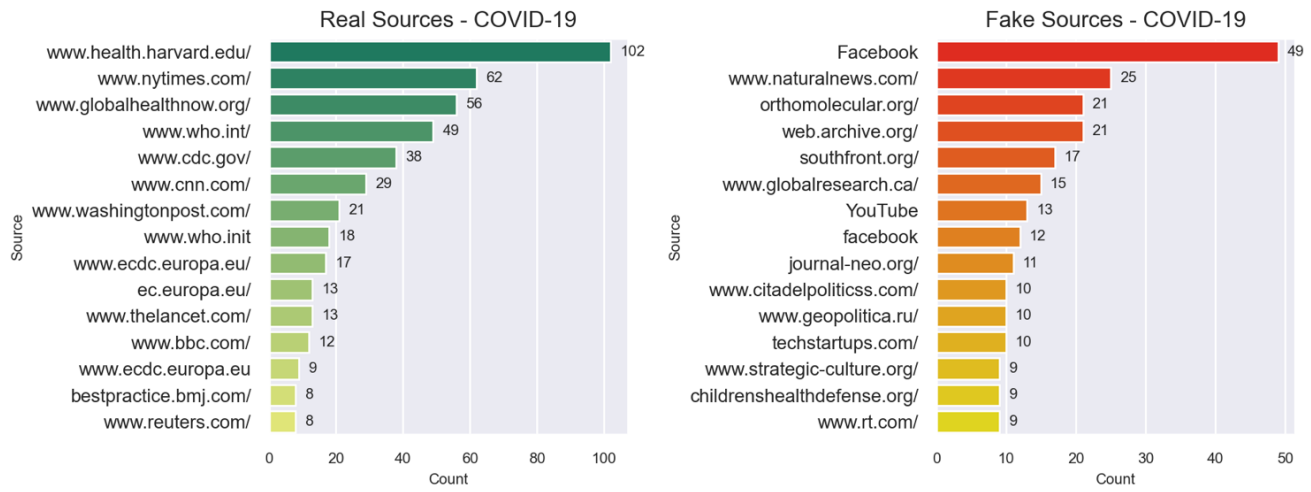


Figure A19: Word Clouds – Real and Fake Text – COVID-19 Articles



A20: Top 15 Real and Fake News Sources – COVID-19 Articles

Modeling – FNID

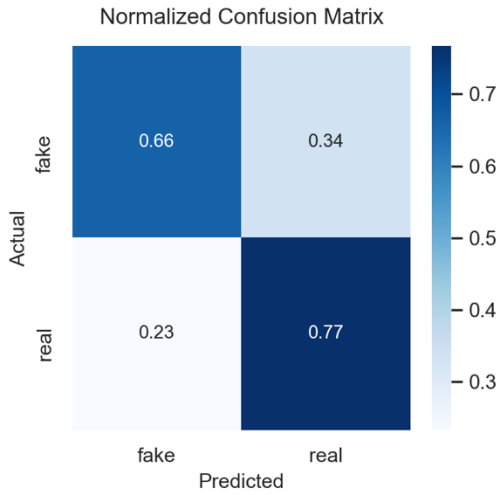


Figure A21: FNID – XGB Normalized Confusion Matrix

Modeling – Combined Topics Dataset

Classifier	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.80	0.805	0.81
Naïve Bayes	0.76	0.79	0.76
Passive Aggressive	0.76	0.76	0.76
Random Forest	0.78	0.79	0.78
XGBoost	0.77	0.77	0.77
RoBERTa	0.78	0.77	0.77

Table A6: Combined Topics Data Fake News Detection Metrics

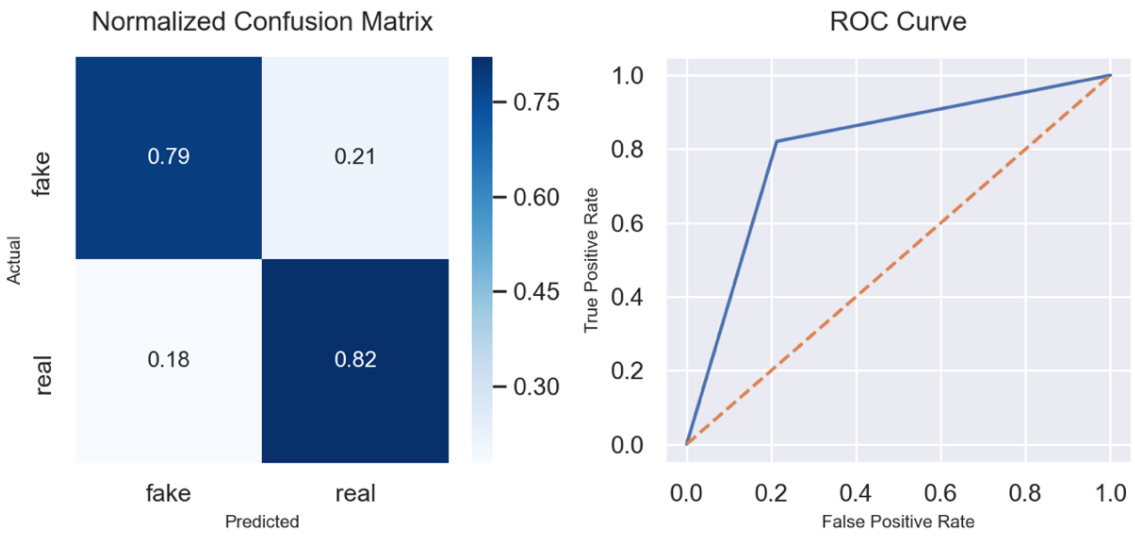


Figure A22: Combined Topics – Logistic Regression Normalized Confusion Matrix and ROC Curve

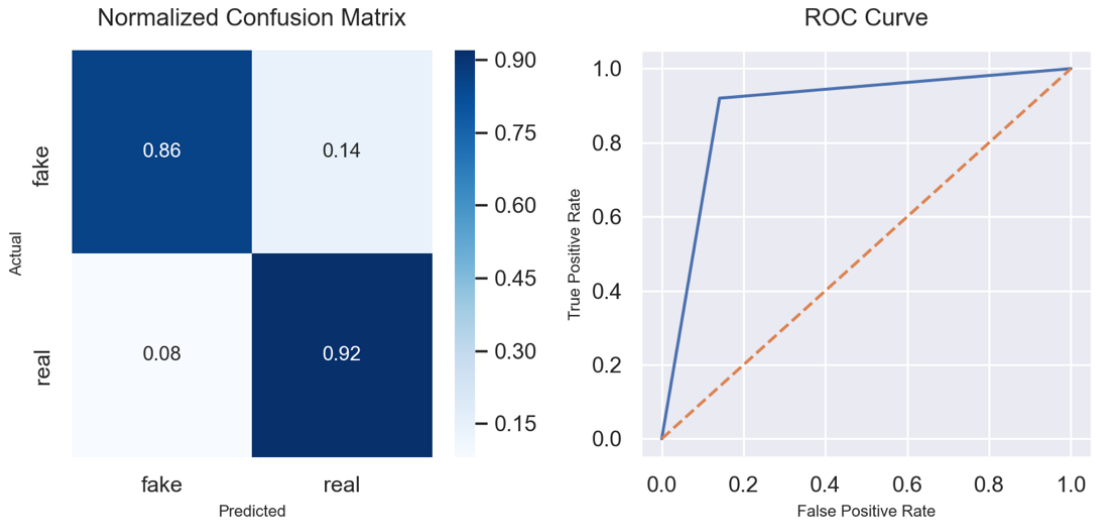


Figure A23: Politics – XGB Normalized Confusion Matrix and ROC Curve

Classifier	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.82	0.82	0.82
Naïve Bayes	0.81	0.805	0.81
Passive Aggressive	0.785	0.79	0.79
Random Forest	0.80	0.81	0.80
Gradient Boosting	0.75	0.77	0.76
XGBoost	0.79	0.80	0.79

Table A7: Gossip Fake News Detection Metrics

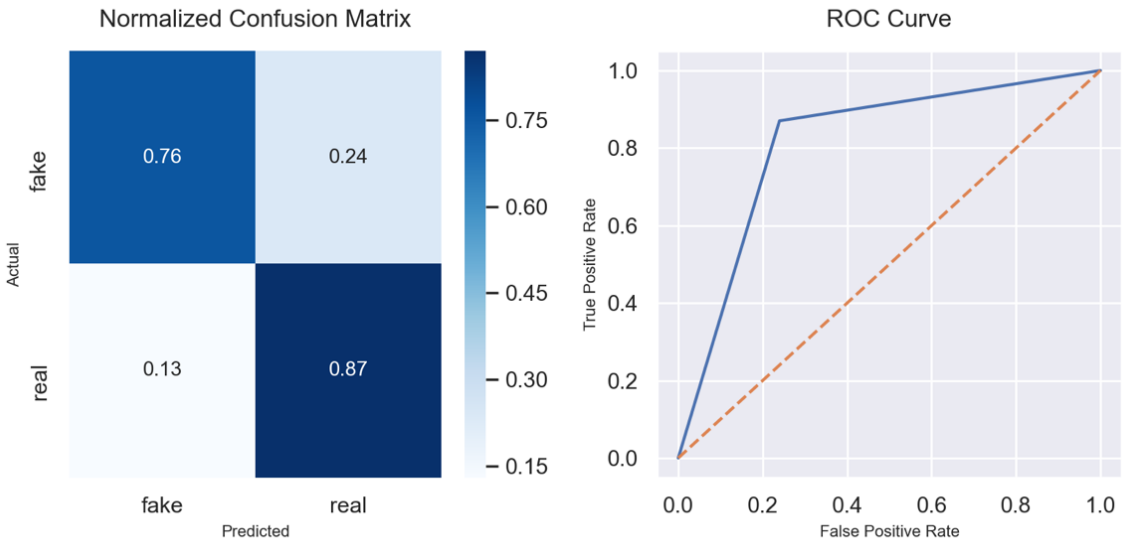


Figure A24: Gossip – Logistic Regression Normalized Confusion Matrix and ROC Curve

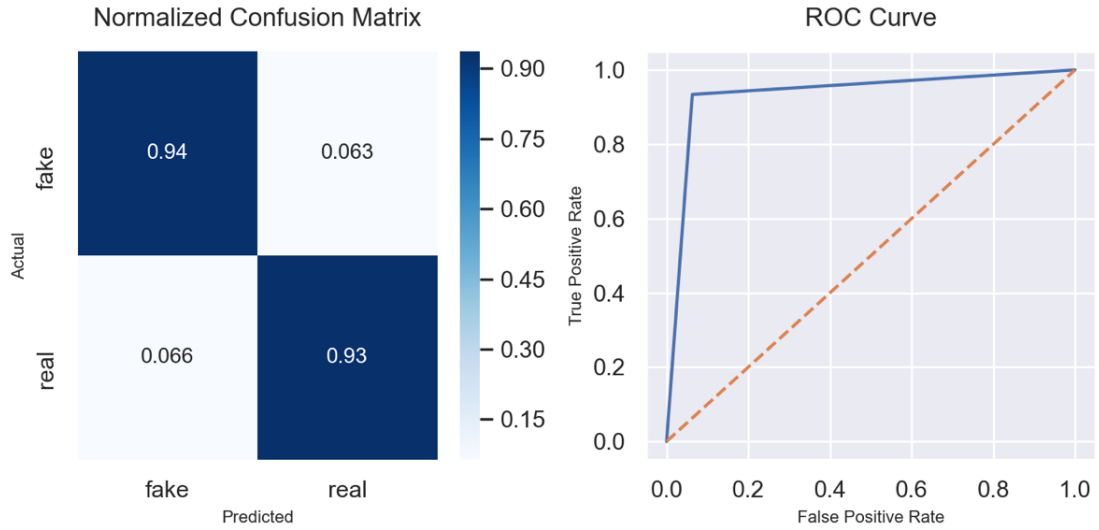


Figure A25: COVID-19 – PA Normalized Confusion Matrix and ROC Curve

Classifier	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.81	0.82	0.81
Naïve Bayes	0.80	0.805	0.80
Passive Aggressive	0.76	0.76	0.76
Random Forest	0.79	0.80	0.79
Gradient Boosting	0.72	0.75	0.72
XGBoost	0.77	0.78	0.77

Table A8: Disasters Fake News Detection Metrics

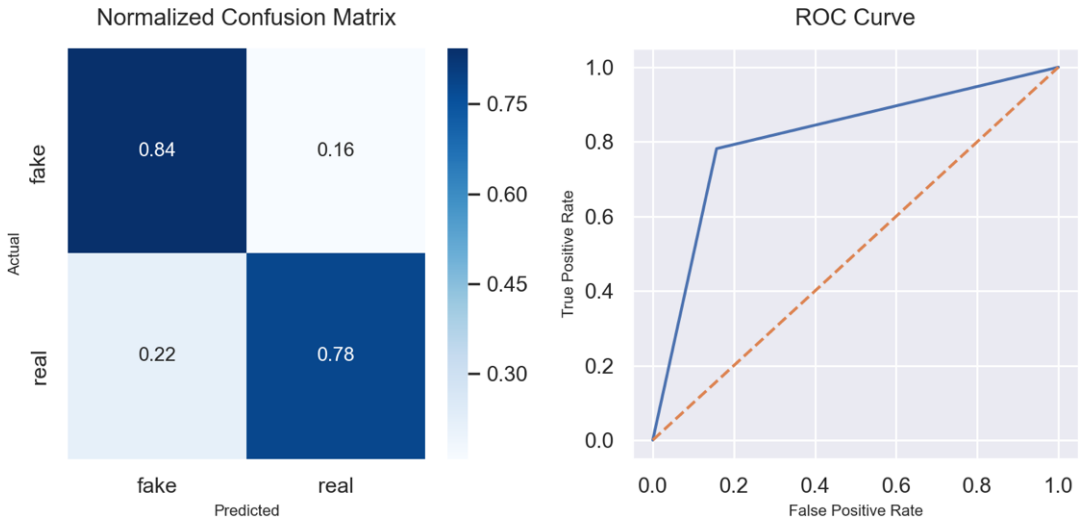


Figure A26: Disasters – Logistic Regression Normalized Confusion Matrix and ROC Curve