Have You Been Fooled?

A Learned Approach to Identify Satire

Team Yavy

Andrew Gruber

Valerie Huang

Yagnesh Patel

Yuqi Wei

With the growing amount of fake news circulating on the internet, the ability to differentiate between real and satirical information becomes crucial. Our group obtained both real and satirical news headlines by either downloading pre-existing datasets from Kaggle, or web scraping from news sites such as the Atlantic and Fox News, as well as satirical websites, including the Onion. We built several machine learning models such as logistic regression, random forest, and neural network to classify whether an article headline was satirical or not. The metrics used to evaluate model performances, such as the AUC, demonstrate that our model indeed has some capacity in distinguishing between authentic and satirical news headlines, although future work is needed to further polish our model and obtain more accurate predictions.

1. Business Understanding

In recent years, significant attention has been paid to the proliferation of 'fake news' on social media[1]. In that conversation, however, the topic of satirical news-media outlets focusing exclusively on satire and making no claims to be truthful represent a grey area. Despite their admitted purpose as sources of entertainment, their linguistic and aesthetic presentations can often lead to misunderstandings. In 2015, global news outlets, including the New York Times, reported that a Pakistani politician claimed that earthquakes were the fault of women wearing jeans[2]. This later turned out to be a piece from The Khabaristan Times, a Pakistani satirical publication. While that may be a gaffe, satirical news masquerading as truth has real implications. In a study evaluating Americans' beliefs in numerous satirical headlines, 53% of those surveyed believed that Disney was moving its Georgia-based operations to California to "avoid filming among depraved, immoral people"[3]. In this way, the inability to determine what is satirical can further deepen political polarization. The same study found that flagging satire was an effective way to communicate the source's credibility to users, with Facebook testing out the

[1] Espaillat, A. (2019, May 03). H.Res.284 - 116th Congress (2019-2020): Opposing fake news and alternative facts. From https://www.congress.gov/bill/116th-congress/house-resolution/284/all-info
[2] Al Arabiya English. (2020, May 20). Media falls for 'jeans cause earthquakes' Pakistan hoax story. From https://english.alarabiya.net/en/variety/2015/05/31/Pakistani-cleric-calls-for-war-against-jeans-wearing-women-.html
[3] Garrett, R., Bond, R., Poulsen, S. (2020, September 03). Too Many People think Satirical News is Real. From https://theconversation.com/too-many-people-think-satirical-news-is-real-121666

feature[3]. Another study shows that 52% of Facebook users rely on the platform to get news[4]. With millions of using social media sites like Facebook to gather their news, an observed risk of falling for satirical news, and a clear demand for the regulation of online platforms, machine learning presents an efficient and automated way to identify satire on social media.

1.1 Related Work

This is not the first exploration into this field of identifying satirical content. Previous work was done with analyzing both the headline and text of articles to determine where they were satirical or fake news[5]. With a class balance of 283 fake news stories to 203 satirical stories, their metrics on their headline only model has a precision 0.46, a recall of 0.89, and an F1-score of 0.78. Other examples had fact-checkers and peers manually flag satirical content, which was deemed effective[6] at deterring misinformation but cannot be done on a large-scale efficiently.

2. Data Understanding

Using web scraping techniques, we obtained data from nine news sources- four satirical (The Onion, Reductress, The National Report, and Hard Times) and five authentic (Fox News, The Atlantic, The Guardian, NPR, and NotTheOnion). NotTheOnion is a subreddit consisting of real headlines that users found to resemble satirical ones. Finally, Kaggle datasets with headlines from ABC Australia and The Huffington Post were also used. By combining data from the 11 different news sources and dropping the duplicates, we obtained a dataset of 439,412 instances and 4 attributes. These attributes are the source, headline, date of the news, and a flag if the

---

[4] Mitchell, A., Jurkowitz, M., Oliphant, J., & Shearer, E. (2020, July 30). Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable. From https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/

[5] Levi, O., Hosseini, P., Diab, M., & Broniatowski, D. (2019). Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. doi:10.18653/v1/d19-5004

[6] *Journal of Computer-Mediated Communication*, Volume 24, Issue 5, September 2019, Pages 240–258, https://doi.org/10.1093/jcmc/zmz012

headline is satirical or not. Figure 1 (Appendix 6.2.1) displays the most common words found in our set of headlines. Certain words are found across all sources- 'Donald Trump', 'USA', 'China'- while others reflect one source's heavy use of a term - 'woman' by Reductress and 'Australian' by ABC Australia.

Our initial EDA helped us uncover the distribution of headlines amongst our sources, the balance of our classes, and spot some oddities within our data. The initial source balance shows that we have collected many more results for some of our sources (Figure 2). While only 5.91% of our data is satirical, 2020 contains a larger proportion of satirical headlines compared to other years (Figure 2). To understand the size of our vocabulary, we have 71,631 unique words, but are able to capture 95% of them with the first 15,720 most popular words. Finally, for models that utilize structured sequences, we looked at the length of headlines, and found a long tail distribution, but had over 99% of headlines falling under 40 words long (Appendix 6.2.2). Among the 6 non-satirical sources, Fox News, The Atlantic, The Guardian and NPR have approximately the same amount of data. On the other hand, ABC Australia and the Huffington Post are overrepresented.

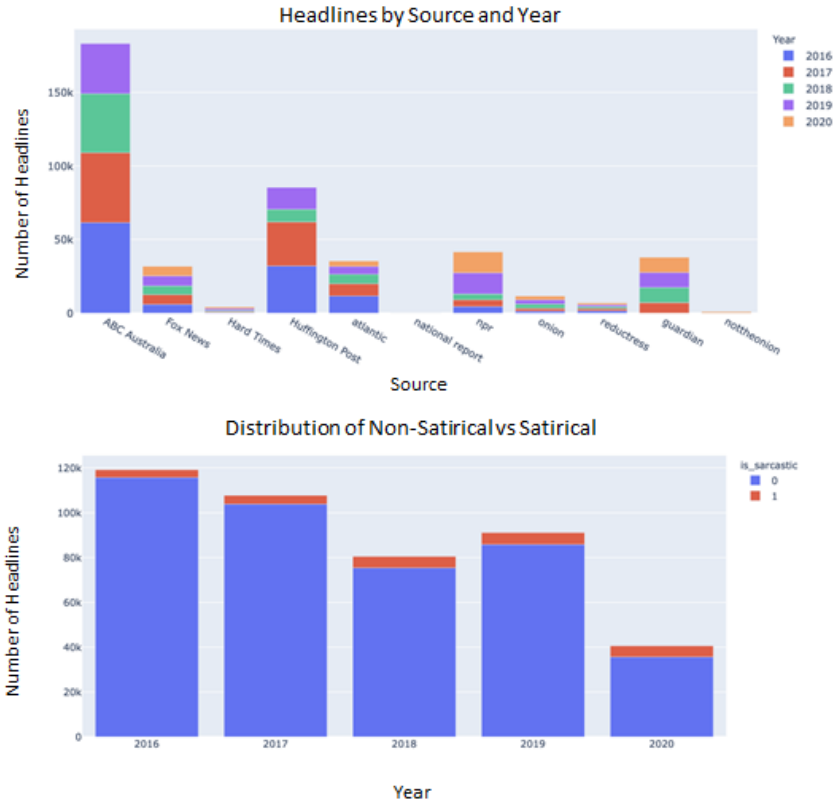


Figure 2: Headlines by Source and Year (top), Distribution of 'is_satirical', by year (bottom)

2.1 Selection Bias

ABC Australia is the only news source that mainly features events outside the US. Due to the large proportion of this source in the combined dataset, headlines from the ABC Australia can be easily identified or predicted as non-satirical

due to their associations with Australia. The opposite bias is present in the US sources, where their US-focused reporting cannot be generalized to global reporting. Another potential source of bias comes from the imbalanced classes. As only 6% of the headlines were satirical, there is concern that our models will over-classify to non-satirical and fail to learn patterns of satire.

3. Data Preparation

For modeling, each instance within our dataset will contain the headline text as the predictor and a binary indicator that classifies each headline as satirical with 1 or not with 0 for the response. We chose to only use the headline to replicate social media users' behavior of only reading the headlines. For our models to better capture language patterns that distinguish satirical content and avoid leaning source specific idiosyncrasies, we decided to down-sample the Huffington Post and ABC Australia, using the Guardian as the baseline. Figure 3 shows the distribution after down-sampling. The 6 non-satirical sources are much more balanced now, with each having approximately 35,000 article headlines. After down-sampling, around 12% of headlines are satirical. We decided to use

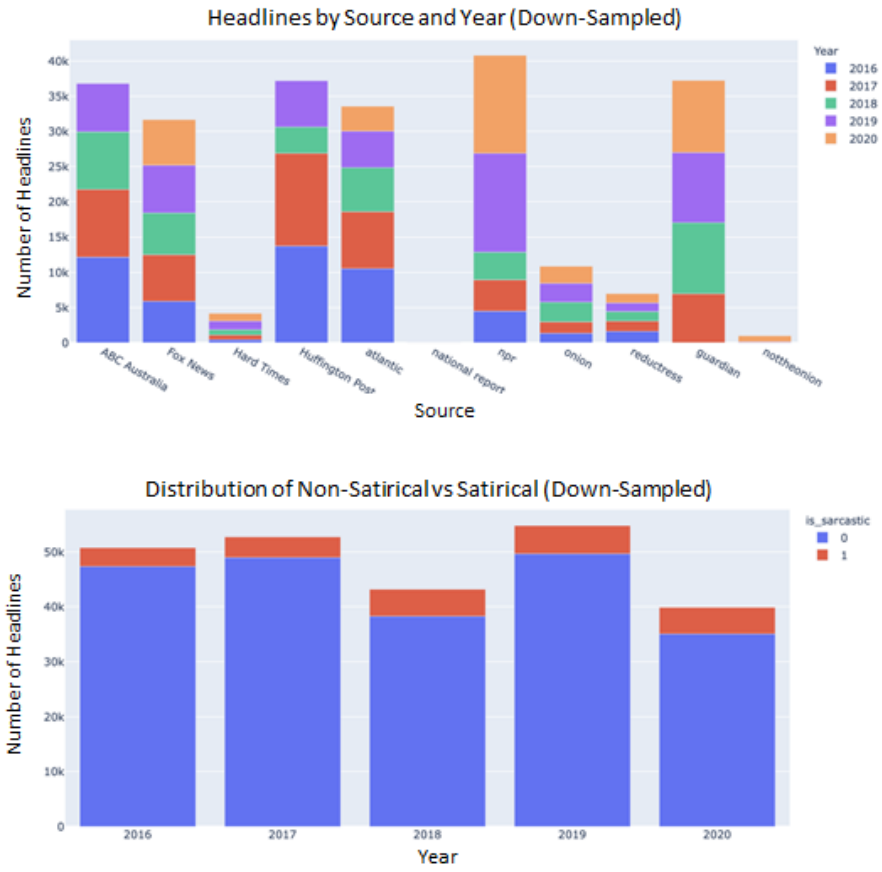


Figure 3: Headlines by Source and Year (top), Distribution of 'is_satirical', by year (bottom) after downsampling

2016-2019 as the training set, and the 2020 headlines as the test set, partly because both 2016 and 2020 are election years, which we assume should share similar news headlines.

3.1 Up-sampling

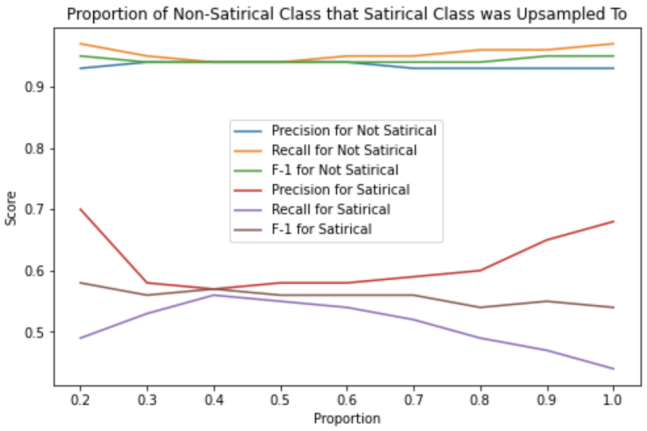


Figure 4: Performance of Different Proportions of Non-Satirical Class to which the Satirical Class was Upsampled

One of our major concerns was the sparsity of satirical headlines. In order to achieve a more balanced data set, we decided to up-sample the minority class. Holding the non-satirical headlines fixed, we up-sampled our train satirical headlines to different proportions of the majority class, ranging from 20% to 100%, when the two classes have equal numbers of headlines. We then fitted these different training sets on a logistic regression model and compared their performances on our fixed test set using precision, recall and F-1 scores. Figure 4 shows the precision, recall and F-1 scores for different proportions of non-satirical headlines. When the number of satirical headlines is 40-60% of the number of authentic ones, the 3 metrics are the most balanced and all generate satisfying results. In order to be more computationally efficient, we up-sampled our satirical data to 40% of the number of non-satirical headlines.

3.2 Special Characters, Lowercase, Contractions and Abbreviations

In order to supply our headline text data to downstream models, we performed common normalization techniques such as lowercasing, expanding contractions, and removing special characters. We also converted common abbreviations that were getting transformed improperly, such as “U.S” becoming “us” to their full form. Additionally, some special characters were converted to their ASCII numeric format and those were also corrected.

3.3 Lemmatization

Lemmatization is the process of reducing words to their base forms (e.g. talks → talk). We explored the three most widely used lemmatizers supported by Python packages: Wordnet NLTK, Spacy, and Gensim[7]**.** Due to different designs, the three lemmatizers yielded slightly different outputs. We ran our models using all three lemmatizers as well as without any lemmatization in order to determine the best performing configurations.

3.4 Profanity and Source Specific Language

The next step in preparation was creating a second version of the data set: one without source specific language and profanity. In our initial test, we found that many of the words associated with a headline being classified as satirical were profane ones. Given that the primary purpose of satirical news is to entertain, it is not a surprise that profane words appeared there much more frequently than in our real news sources. In order to prevent the model from only learning that profanity as an indicator of satire, the alternate data set excluded all profanity and all source specific language. By source specific language, we refer to certain words or phrases used nearly exclusively by one source. *The Onion*, for example, publishes horoscopes weekly leading to many headlines reading "Horoscope of the Week". This sort of filtering was done for all sources and any headline incorporating a source's specific section (e.g. *The Atlantic*'s 'Poem of the Week') or terms used by predominantly one source (e.g., *ABC Australia*'s use of 'QLD' to refer to Queensland) were excluded from this data set.

4. Modeling & Evaluation

4.1 Evaluation Metrics

We used AUC as our primary method to compare model performance. It is desirable in

[7] Prabhakaran, S. (2020, May 18). Lemmatization Approaches with Examples in Python. From https://www.machinelearningplus.com/nlp/lemmatization-examples-python/

our situation as it is class balance invariant. However, we also compared precision, recall, and F1-score to measure the performance within the two classes. We believe that both precision and recall are critical, as from the perspective of the deployment companies, misclassifying real news as satirical harms their partnerships with reliable news sources; misclassifying satirical as real can also incur severe consequences such as the spread of misinformation and the lowered trust in social media. Such consideration makes F1-score another good metric to use since it is a balanced summary of precision and recall.

4.2 Baseline - Logistic Regression

We used logistic regression as our baseline model because it is robust under imbalanced classes, easy to interpret, and inexpensive to run. We first used the original data that includes profane words and source-specific language to tune hyperparameters. The cleaned text was fed into a vectorizer (CountVectorizer or TFIDF vectorizer, with TFIDF performing better) that returned binary outputs of unigrams, bigrams, and trigrams. By utilizing a 5-fold GridSearchCV, we were able to get the best performing configuration with the complexity parameter C = 10, an L2 penalty, and maximum number of iterations = 1000. The AUC score for predicting 2020 headlines was 0.88, which was surprisingly high. Looking at the top features (Appendix 6.3.1-2), source and geographical specific words were crucial in predicting real headlines, while profane words played a key role in predicting satirical ones. Such a discovery reinforced the need to remove those words for better generalizability. The logistic regression with non-source data gave a slightly lower AUC of 0.87, still indicating good performance for a baseline model.

4.3 Additional Models

Hoping to improve from the baseline, we experimented with models that worked well with sklearn's sparse matrix format and were suited for classification tasks. This led us to

experimenting with Random Forest, Naive Bayes Models, and Support Vector Machines. For Random Forest, we used cross-validation to learn the optimal hyperparameters n_estimator = 70 and min_samples_leaf = 2, as suggested by grid search. However, with an AUC of 0.84 and a recall of 0.46 for the original dataset, and an AUC of 0.83 and a recall of 0.45 for the reduced, random forest failed to surpass logistic regression. We suspect that random forest's poor performance can be attributed to the extremely sparse data, which caused the algorithm to build trees with useless words.

Two versions of the Naive Bayes, Bernoulli and Multinomial, were fitted on the original data set and the reduced version without profanity and source specific language. Using GridSearchCV, the smoothing parameter, alpha, was tuned to be 1 in both Multinomial Naive Bayes models and tuned to be 5 and 4 in the Bernoulli Naive Bayes models in the full and reduced data sets, respectively. While the Naive Bayes models were outperformed in precision and recall on the non-satirical headlines, their precision and recall for the satirical headlines were among the best (See Page 10).

We initially tried Support Vector Machines but found the training time to be too slow. Due to limited computing power, we ran the default SVM model and discovered the results were inferior to the baseline logistic regression and excluded SVMs from further evaluation.

4.4 Neural Networks

Neural Networks are a common modeling approach for NLP tasks as they are able to take advantage of multiple levels of representations such as taking into account the ordering of sequences and discovering relationships (embeddings) between words[8]. During our experimentation, we tried several different architectures. In general, we sent the tokenized

[8] Di, W., Bhardwaj, A., & Wei, J. (2018, April). Why Deep Learning is perfect for NLP (Natural Language Processing). From https://www.kdnuggets.com/2018/04/why-deep-learning-perfect-nlp-natural-language-processing.html

headlines padded/limited to a length of 40 into an embedding layer and then through either convolutional or RNN layers before being passed through several dense layers to get a prediction outcome. Convolutional and RNN layers can learn the sequential nature of the words by learning not only what is important, but when it is important. For the embedding layer, we tried both to learn the embeddings and to use pre-trained GloVe[9] embeddings. Because we tried pre-trained embeddings, we did not use a lemmatizer in these models.

Using an 80-20 train-validation split, the best performing model we obtained utilized self-learned embeddings and Bi-Directional LSTMs with a hidden size of 256 (Appendix 6.2.4). For our source language-included test set, this model achieved an AUC of 0.91 and an F1-score of 0.63. On our non-source dataset, we obtained an AUC of 0.91 and an F1-score of 0.60. We saw quick convergence when it came to epochs, with our optimal validation AUC often occurring between 3 to 5 epochs (Appendix 6.2.3). While we anticipated that the GloVe embeddings would outperform our self-learned embeddings, the pre-trained embeddings may have fallen short due to missing 5% of our vocabulary and words in a headline differing in context from general representations.

4.5 Model Comparison

Source-Language Included

Model	AUC	Precision	Recall	F1-Score
Neural Network	0.9159	0.94 (0) 0.83 (1)	0.99 (0) 0.51 (1)	0.94 (0) 0.63 (1)
Bernoulli Naïve Bayes	0.8893	0.94 (0) 0.61 (1)	0.95 (0) 0.58 (1)	0.95 (0) 0.61 (1)
Multinomial Naïve Bayes	0.8886	0.94 (0) 0.70 (1)	0.97 (0) 0.52 (1)	0.95 (0) 0.60 (1)
Logistic Regression	0.8761	0.94 (0) 0.60 (1)	0.95 (0) 0.58 (1)	0.90 (0) 0.59 (1)
Random Forest	0.8413	0.93 (0) 0.62 (1)	0.96 (0) 0.46 (1)	0.94 (0) 0.53 (1)

Source-Language Excluded

Model	AUC	Precision	Recall	F1-Score
Neural Network	0.9172	0.93 (0) 0.85 (1)	0.98 (0) 0.46 (1)	0.94 (0) 0.59 (1)
Bernoulli Naïve Bayes	0.883	0.94 (0) 0.61 (1)	0.95 (0) 0.58 (1)	0.94 (0) 0.59 (1)
Multinomial Naïve Bayes	0.8784	0.95 (0) 0.55 (1)	0.93 (0) 0.64 (1)	0.95 (0) 0.59 (1)
Logistic Regression	0.8676	0.94 (0) 0.58 (1)	0.95 (0) 0.55 (1)	0.90 (0) 0.56 (1)
Random Forest	0.8317	0.93 (0) 0.59 (1)	0.96 (0) 0.45 (1)	0.94 (0) 0.50 (1)

[9] GloVe: Global Vectors for Word Representation https://nlp.stanford.edu/projects/glove/

After acquiring the best hyperparameter choices for each of the models we used, the neural network is nearly universally the best performer. Whether source language was included or not, it achieved an AUC over 0.91. Additionally, its positive class F1-score at a 0.5 threshold is either better or tied with the other models. Upon further review of the ROC curves and precision-recall curves, the neural network's curves dominate the others (Appendix 6.2.5). While the LSTM did perform very well, one clear drawback is that the model is the most complex option and takes approximately 2-5 times longer to train compared to the others. Despite this, we believe that the performance warrants the additional complexity and would go with the LSTM.

5. Deployment

The implementation of this classification algorithm would include collaboration with social media companies like Facebook and Twitter to add a feature that identifies content as satirical. One obstacle we anticipate is that platforms might hesitate about investing in such a classification feature, given that eye-catching satirical headlines might attract more clicks, comments, and reposts. However, the public would be well served by social media platforms that distinguish satire from truth, especially during special periods such as the election when satirical news surges. As many platforms already have fake news warnings, adding satire flags would not be a difficult endeavor. Our classifier would be particularly useful in identifying the headlines of lesser-known news sources. Presumably, the companies that deploy our classifier would always mark The New York Times as non-satirical. This model would be able to classify headlines of unknown or obscure origin. By navigating users through an increasingly complex news environment, our classifier brings more benefits than costs.

5.1 Continual Monitoring & Training

In order to prevent the model from becoming errant or growing stale, we propose the following strategy. To gather new data, we would subscribe to well-established satire and non-satirical RSS news feeds to bring in daily headlines and also allow us to automatically tag each headline's class based on source. This would help us grow our dataset, maintain a consistently fresh dataset, and be less error prone than trying to build custom web-scrapers. With this fresh supply of classified headlines, we can regularly retrain our model with the latest data to avoid concept drift. We can also use this process to monitor if our model performance significantly drops or to observe any faulty RSS feeds by a lack of new content from a source.

5.2 Ethical Considerations

There are several ethical considerations to keep in mind, on both the model deployment side and the user interface side. The misclassification of real headlines could harm the reputation of actual media outlets and prevent the spread of accurate information. The misclassification of satirical headlines would allow for the spread of misinformation on social media platforms. Language is specific to publications and individual writers making perfect classification a pipedream. While our model has learned the words and phrases most strongly associated with satirical writing, it cannot always ensure that satirical writing will be classified as such, nor can it ensure the quality of journalism from real news headlines.

6. Appendix

6.1 Bibliography

Al Arabiya English. (2020, May 20). Media falls for 'jeans cause earthquakes' Pakistan hoax story. From https://english.alarabiya.net/en/variety/2015/05/31/Pakistani-cleric-calls-for-war-against-jeans-wearing-women-.html

Di, W., Bhardwaj, A., & Wei, J. (2018, April). Why Deep Learning is perfect for NLP (Natural Language Processing). From https://www.kdnuggets.com/2018/04/why-deep-learning-perfect-nlp-natural-language-processing.html

Espaillat, A. (2019, May 03). H.Res.284 - 116th Congress (2019-2020): Opposing fake news and alternative facts. From https://www.congress.gov/bill/116th-congress/house-resolution/284/all-info

Garrett, R., Bond, R., Poulsen, S. (2020, September 03). Too Many People think Satirical News is Real. From https://theconversation.com/too-many-people-think-satirical-news-is-real-121666

GloVe: Global Vectors for Word Representation https://nlp.stanford.edu/projects/glove/

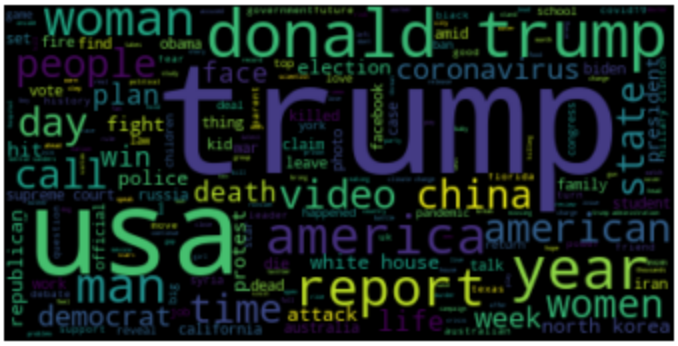
Journal of Computer-Mediated Communication, Volume 24, Issue 5, September 2019, Pages 240–258, https://doi.org/10.1093/jcmc/zmz012

Levi, O., Hosseini, P., Diab, M., & Broniatowski, D. (2019). Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. doi:10.18653/v1/d19-5004

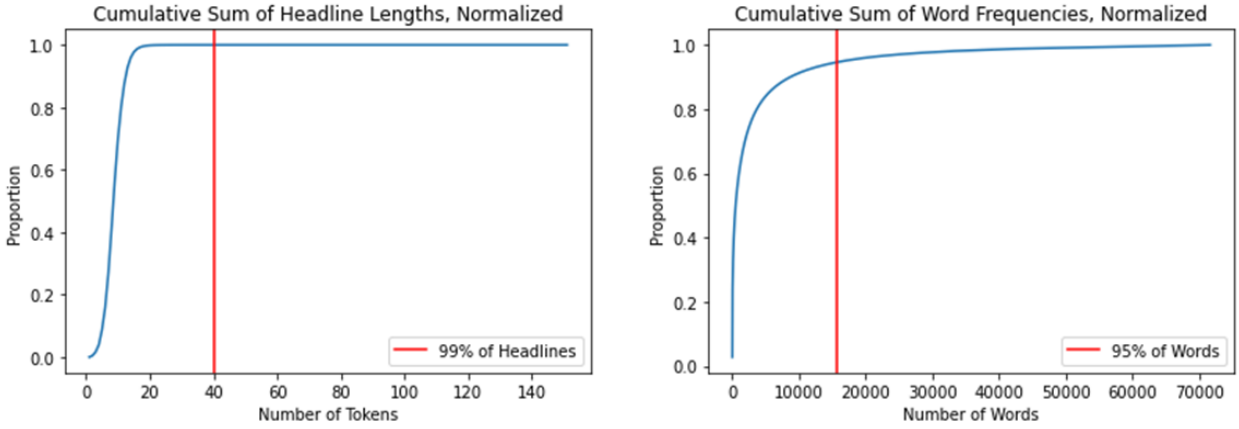
Mitchell, A., Jurkowitz, M., Oliphant, J., & Shearer, E. (2020, July 30). Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable. From https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/

Prabhakaran, S. (2020, May 18). Lemmatization Approaches with Examples in Python. From https://www.machinelearningplus.com/nlp/lemmatization-examples-python/

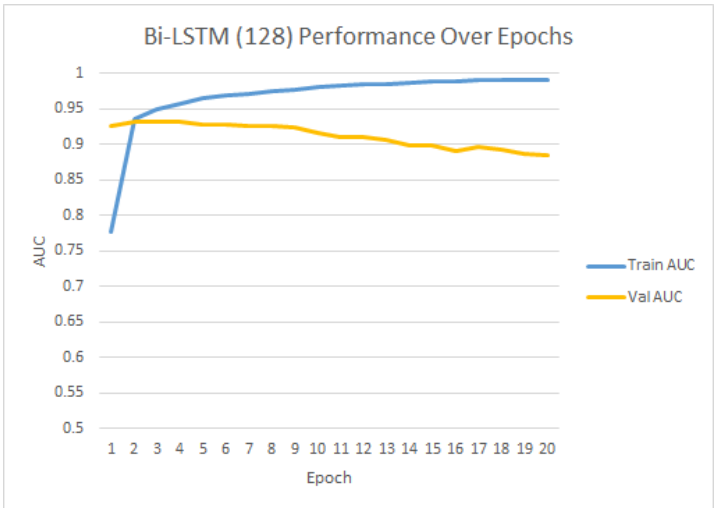
6.2 Figures



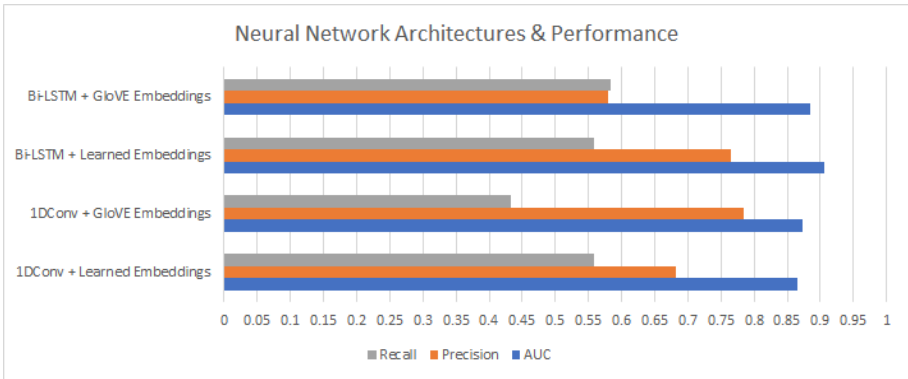
6.2.1: Word cloud of most common words and phrases in the complete data set, with stopwords excluded.



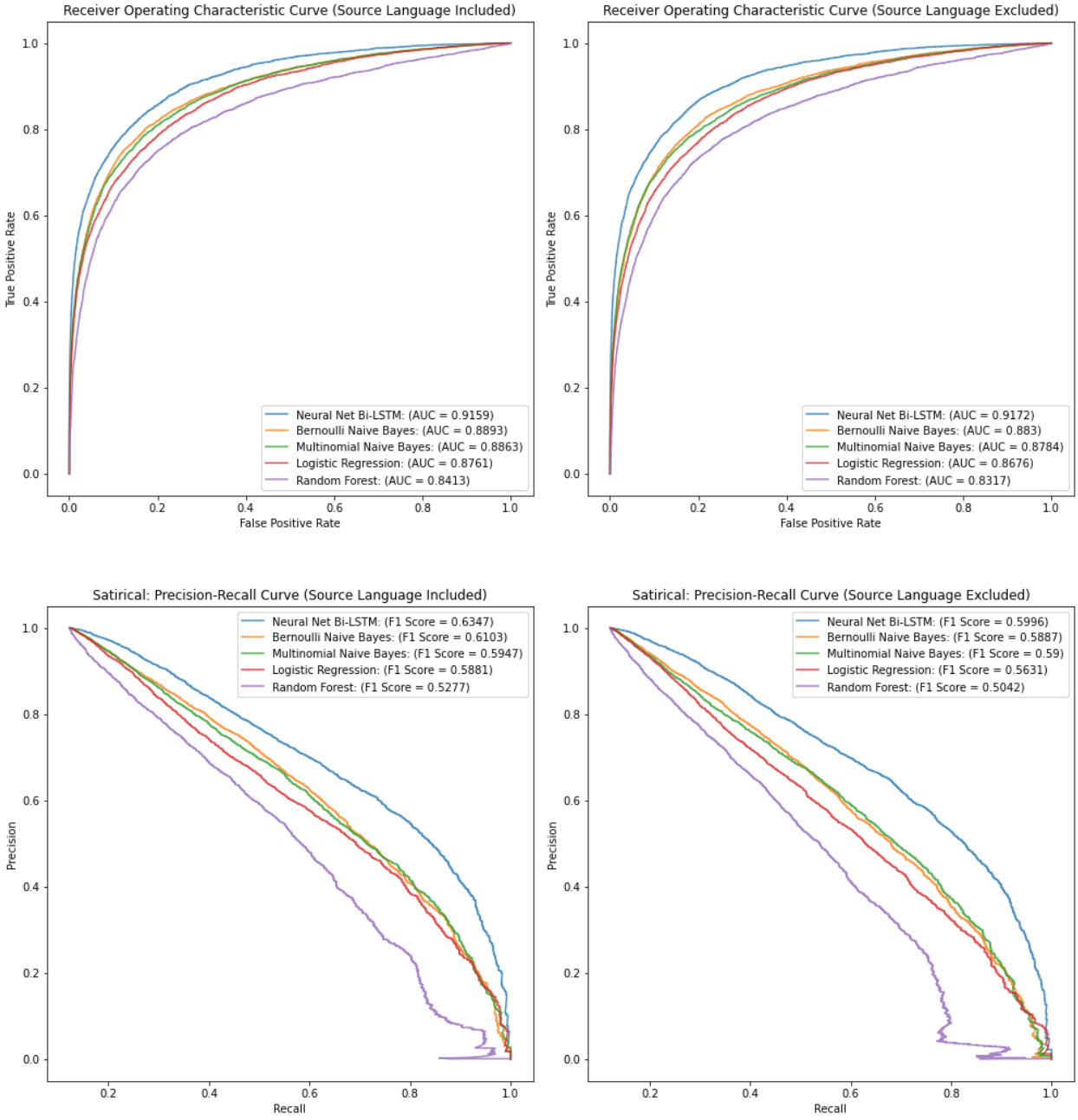
6.2.2 Cumulative Distributions of Headline Lengths and Word Frequencies



6.2.3 Performance of Sample Neural Network Model over Epochs

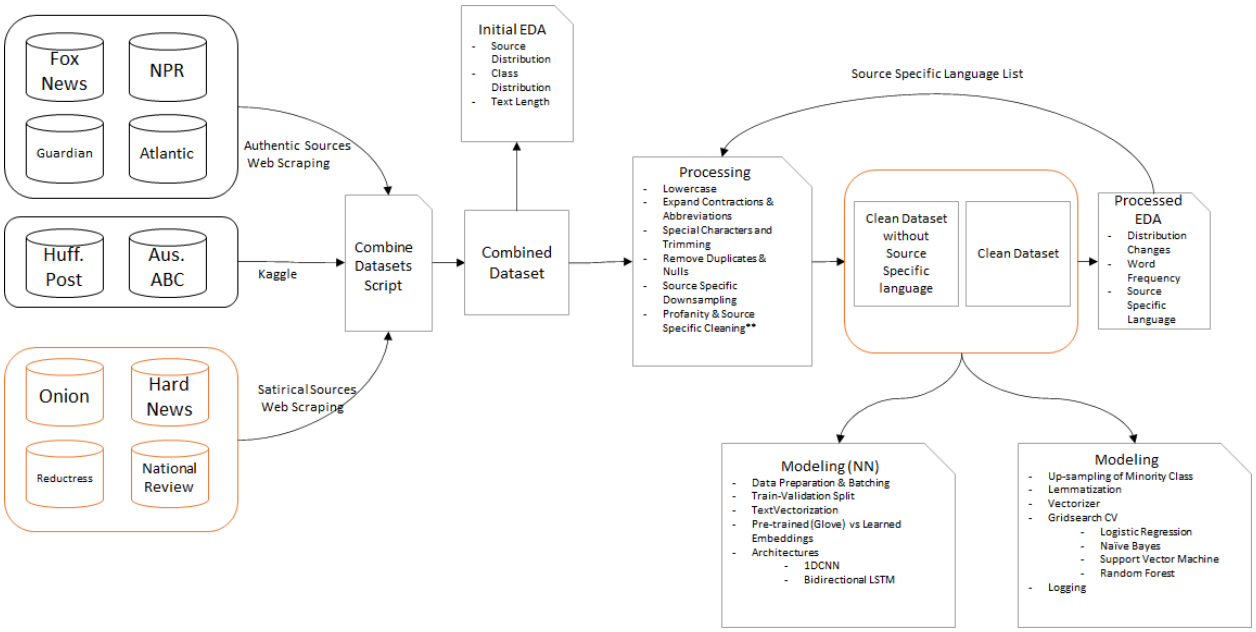


6.2.4 Model Performance of Various Neural Network Architectures



6.2.5 Performance chart of final models after tuning

6.2.6 Development Flow



6.3 Tables

Words with Lowest Coefficients of LR Model (have most predicting power for real headlines)	
Rank	*Words*
1	report says
2	melbourne
3	hobart
4	wa
5	track day
6	adelaide
7	sydney
8	canberra
9	dear therapist
10	surprising

Words with highest Coefficients of LR Model (have most predicting power for sartirical headlines)	
Rank	*Words*
1	fucking
2	punk
3	trump boys
4	onion
5	wow
6	fuck
7	shit
8	things know
9	assures
10	horoscopes

Table 1. Top Features BEFORE Removing Profanity and Source-Specific Language

Words with Lowest Coefficients of LR Model (have most predicting power for real headlines)	
Rank	*Words*
1	report says
2	amid
3	alleged
4	rio
5	aged
6	percent
7	study says
8	colbert
9	pennsylvania
10	viral

Words with Highest Coefficients of LR Model (have most predicting power for sartirical headlines)	
Rank	*Words*
1	punk
2	trump boys
3	wow
4	realizing
5	informs
6	assures
7	gamer
8	band
9	english teacher
10	clinging

Table 2. Top Features AFTER Removing Profanity and Source-Specific Language

6.4 Contribution

Andrew Gruber: Bernoulli/Multinomial Naive Bayes, Data Scraping (Reductress, The Atlantic, The National Report), Source Specific Language and Profanity, Research, Ethical Considerations

Valerie Huang: Random Forest, Data Cleaning and Combination, Up-sampling and Down-sampling, Visualizations, Abstract

Yagnesh Patel: Neural Networks, Data Scraping (Fox, The Onion, Hard News, NotTheOnion), Related Work Research, Visualizations, Continual Training, Development Flow

Yuqi Wei: Logistic Regression, SVM, Data Scraping (Guardian, NPR), Lemmatization (Spacy, Gensim), Evaluation Metrics Comparison, Feature Ranking, Deployment

Overall: Business Problem Research, Data Gathering and Exploration, Modeling Approaches, Reporting Writing, Proofreading