COMP90042 Project 2020: Climate Change Misinformation Detection

Haripriya Ramesh
Student ID: 1073646
rameshh@student.unimelb.edu.au

Abstract

There is a constant menace caused by fake news spreading through social media outlets. One such topic is the spreading of misinformation pertaining to climate change. This is often aimed at misleading people and can have serious impacts. There is an increasing need to develop a system that can detect and classify such news as fake. In this project, I attempt to detect if an article contains climate change misinformation with the help of Natural Language Processing techniques.

1 Introduction

The main aim of the project is to identify climate change misinformation. For this task, I was provided with training data that contains 1169 sample articles that have misinformation regarding climate change. There are multiple approaches to tackle this problem. It can be viewed as a one class classification task where we attempt to detect the outliers by using the provided positive label data as supervision. Another approach is to expand the training data to include documents with real news that will have negative label and build a binary supervised classifier.

For a comparative study, I explored the following approaches:

- Outlier detection with One-class SVM.
- Doc2Vec transformation and similarity threshold provided training data and classification based on similarity threshold.
- Binary classification with Deep learning.
- Binary Classification with logistic regression.

The overall best performance was given by logistic regression with CountVectorizer after tuning the hyperparameters with the final F1 score of 56 on the provided test data. In this report I have provided a brief analysis of outlier detection with OneClassSVM, binary classification with deep learning and binary classification with Logistic Regression, followed by evaluation of results, error analysis and finally the future enhancements and conclusion.

2 Methodology

I conducted multiple experiments using various approaches ranging from one class classifier to deep learning all using existing python programming libraries. For outlier detection I used oneClassSVMs and unsupervised deep learning using doc2vec tranformation on text with auto-encoder. I further conducted experiments by extending the training set to include real news and performed binary classification using Logistic Regression, Naive Bayes and a few other techniques. Additionally deep learning was also attempted but it did not give satisfying results due to the improper architecture.The development of the models were done in stages where preliminary analysis was done without performing hyperparameter tuning. Further improvement was done by analysis of the scores obtained with the validation set and feature engineering. Grid Search with cross validation was also used to perform hyperparameter tuning. The approaches are further briefed below.

2.1 Dataset

The dataset provided for this project has 1168 training articles that belong to the misinformation class, 100 samples for validation with labels indicating real or fake news and test data with 1410 articles to be classified. Further ’All the news’ dataset was taken from kaggle which consists of 143,000 articles from 15 American publications and has many columns including ’title’,’publication’, ’author’, ’date’ and ’content’(Kaggle). To extend the task to perform binary classification, I took this data set and took 1168 random samples from the set to account for the real news in the training data. Upon investigation, further tweaking was done to include articles with climate change content by pattern matching with the title column to extract rows that have the pattern ’climate change— global warming. 100 such articles were found and taken along with 1068 random samples. The top Unigrams in the provided fake news (climate change misinformation) set can be seen in Figure 3 and those in real news can be seen in Figure 2.

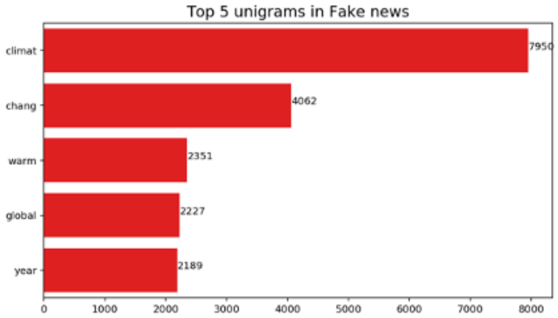

Figure 1: Unigrams in Fake news

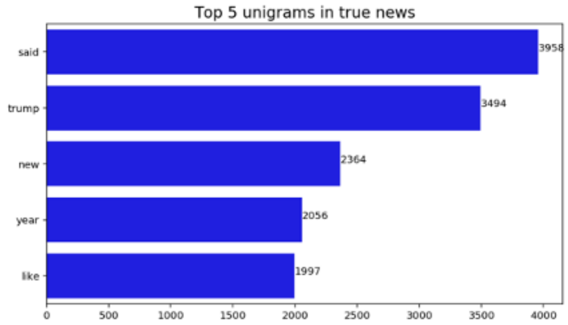

Figure 2: Unigrams in Real news

2.2 Outlier Detection with One Class SVM

One-Class SVM is a supervised algorithm that is mainly used in novelty detection. This can be used to classify new data as similar or different to the training set. This fits our task as climate change misinformation detection with the provided training set can be modeled using One-Class SVM.

2.2.1 Pre-processing and Doc2Vec Transformation

Most news articles tend to have ellipsis, question marks, and special characters which can be removed. As with any text classification task, preprocessing was performed by removing special characters, stopwords, and also stemmed using gensim package tools. Further for this task, for converting the text to vector space model, Doc2Vec transformation was used. The main motivation of using Doc2vec was that it is very good in understanding well written text such as news information.

2.2.2 Classification with oneClass SVM

The positive class in this project is the class encompassing climate change misinformation data. One-class SVM works by learning the boundary of the positive class and for any new point, if it falls inside the boundary it considers it be positive. All other points outside of that point are classified as negative, i.e., real news. Grid Search was performed to tune the hyperparameters of the classifier. This model got an F1 score of 68 on the validation data and F1 score of 39 on the test data. The ROC curve of the model is shown below in Figure 3.

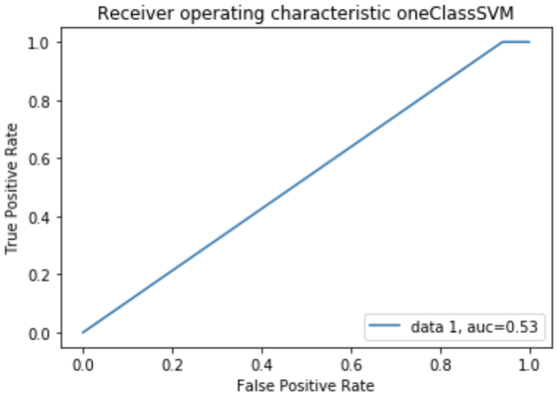

Figure 3: ROC One-ClassSVM

2.3 Binary Classification using Deep Learning

Motivated by many state of the art approaches that use neural networks to approach this problem of telling fake news apart from real news, I conducted similar experiments in this project. This architecture of the Neural Network was adapted from (Wang, 2017). Further GloVe file was used for generating word embeddings.

The neural network implemented consisted of various layers. The first layer is a convolution layer which was use to extract features from the data and improve performance followed by max pooling to take the highest value tensors and to compress the features. Next, LSTM layer was used which will help remember values over time. Dropout is also used to prevent the model from over fitting. At the last layer, sigmoid activation function is used as we have a binary classification task.

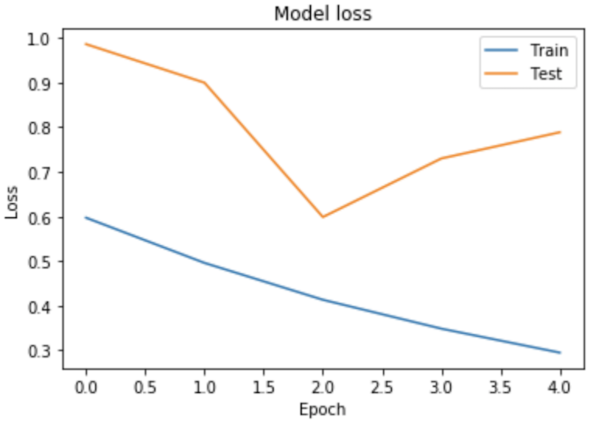

Figure 4: Loss Curve for 5 epochs with Validation set and training set

The From Figure 4 we can say that the model does not have sufficient capacity. Even after increasing the number of epochs the model showed similar behavior which goes to show that the architecture of the model needs to be modified. We can say that from the development set loss curve the model is not generalising well.

2.4 Binary Classification with Logistic Regression

To further improve the performance of my model, I shifted focus to simpler binary classification models. The training data was expanded to include samples from real news as described in the dataset section. My aim was to start with basic classification models, analyse the performance and improve it(Granik and Mesyura, 2017). Exploratory Data Analysis was done to identify of most frequently used bigrams and trigrams to prepare a custom stopwords list.

A combinations of vectorizers and classification models was considered and using Grid Search Cross Validation the model with the best parameters was considered. Many classifiers including MultinomialNB, XGBoost and SGDClassifier were considered. The best score was given by CountVectorizer with ngram range of 1 and the Logistic Regression combination.

2.4.1 Pre-processing and Feature Analysis

Before the text pre-processing, feature analysis was performed to discover interesting patterns in training data which can help the classifier improve accuracy. Features such as number of questions, ellipsis, uppercase letters used proved to be good features for this classification task. Using these as features in addition to CountVectorizer to transform the text helped improve the F1 score in the Validation set from 77 to 80. We can see in figure 5 how the feature number of questions separates the data in training set.

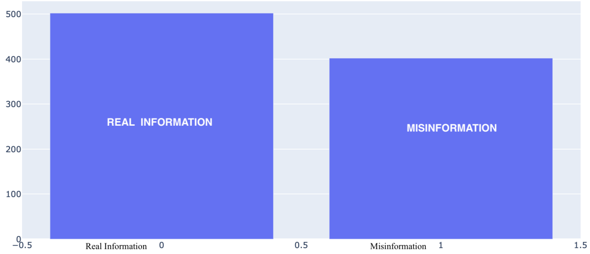

Figure 5: How the feature: Number of Question separates the articles in training data

2.4.2 Classification with Logistic Regression and CountVectorizer

A combination of ngram ranges between 1 and 3 were also studied. The best performing classifier was given by CountVectorizer with ngram range of 1 and the logistic regression combination which was obtained using the validation set. Attempts were made to analyse the coefficients to see how the predictions are made. We can see from Figure 6 that 'climate', 'energy', and 'green' contributes the most to being from fake news class/ misinformation class where as 'it', 'said' and 'trump' indicate that it is likely real news.

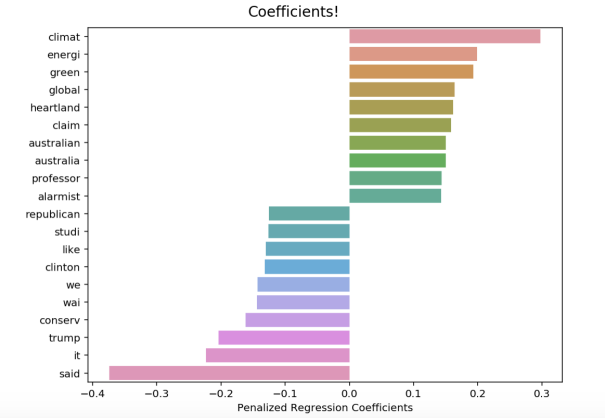

Figure 6: Logistic Regression Coefficients

3 Evaluation

Table 1 shows the consolidated results of all the approaches used. The logistic regression approach

One-ClassSVM

	F1	Precision	Recall
Dev Set	0.662	0.5	0.98
Test Set	0.398	-	-

Neural Network Binary Classification

	F1	Precision	Recall
Dev Set	0.77	0.72	0.84
Test Set	0.5029	-	-

Logistic Regression Binary Classification

	F1	Precision	Recall
Dev Set	0.787	0.66	0.96
Test Set	0.58	0.42	0.92

Table 1: Classification Report of approaches used

got the highest F1 score. This can be seen as the coefficients of the model have clearly indicated the importance of each variable and as the training set has more climate related information that is miss-classified, it performs better than other approaches. The deep learning model considered has a very complicated architecture that needs more fine tuning. One class SVM has given an F1 of 39 on test set due to the size of the training set and lack of efficient feature engineering.

4 Error Analysis and Key Observations

Few key observations and error analysis are listed below.

1. We can see from all three approaches used that recall is high. This goes on to show that out of all the articles identified as fake news, most of them are indeed fake this can be attributed to the presence of key differentiating terms like ’climate change’ in the dataset.

2. Another observation is that the development set provided is not representative of the training set used to train the model. This results in poor F1 scores when testing.

3. Upon analysis of the development set results, it was found that most of articles that were miss-classified as fake news, has the frequent unigrams and bigrams.

4. The neural network approach failed to give results as expected mainly because of poor architecture of the model.

5. There is a large gap between the scores seen in development set as compared to test set mainly because of the lack of training samples. The test articles of 1410 articles are more than training samples provided, giving rise to inconsistent results.

6. Feature engineering by including additional features such as number of questions, exclamations, uppercase letters, etc showed improvement of 1-2 scores increase in performance in the validation set. but when trying with the test set showed no improvements or further performed poorly. This is mainly because the chosen extended training set does not represent the test set.

5 Enhancements

Multiple ways of enhancing the chosen approaches are possible. One such improvement area is the choice of the extended training set. The real news set chosen should be more representative of the text in testing set. Further we can improve performance by a great extent by choosing richer features like POS tags. This way we can capture the semantic intuition of text rather than syntactic like we do here with logistic regression.

Further we can use a hybrid model of different classifiers because some may be better than the others at capturing specific features. Lastly we can use state of the art models like ELMo or BERT which will give better performance (Yang et al., 2019).

6 Conclusion

The task of climate change misinformation detection is analysed using three approaches. The logistic regression classifier and countVectorizer with an ngram range of 1 gives the highest F1 score of 56 in this task. There is a lot of scope for improvement in the current approaches which can be done by using richer features or other state of the art models to capture better the semantics of the text.

References

M. Granik and V. Mesyura. 2017. Fake news detection using naive bayes classifier. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pages 900–903.

Kaggle. All the news.

William Yang Wang. 2017. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. *CoRR*, abs/1705.00648.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fake news detection as natural language inference. *CoRR*, abs/1907.07347.