Climate Change Misinformation Detection

Maria Viji Rashmi
Student Id: 1040446

Abstract

The purpose of the project was to detect if a document contains climate change misinformation. The project was approached by leveraging various feature engineering and data modelling techniques. The results of each technique were compared and contrasted with each other. They were evaluated based on precision, recall and F1 scores. Support Vector Classifier (SVC) performed the best in detecting climate change misinformation, with an F-score of 0.75 on test set. Since it generalized well on test data, it was chosen as the final model.

1 Introduction

Climate change misinformation is a serious issue that requires remediation. This project aims to leverage NLP and statistical modelling techniques to identify misinformation. The project was dissected into 4 phases. These were, data collection, feature engineering, data modelling and evaluation.

2 Methodology

This section describes the methodology adopted in each of the 4 phases.

2.1 Data Collection

The negative class was extended with news articles from various categories such as sports, current events and entertainment. (bbc),(cnn),(gua),(abc) To further improve the precision of the model, climate change news from reliable sources such as NASA and other government websites were added to the negative class dataset.(nas),(nat),(car)

2.2 Feature Engineering

Feature engineering involved selection and extraction of text by leveraging various feature extraction methods. The following describes text pre-processing, feature selection and extraction methods utilized.

- Text Pre-processing: This involved tokenization of articles, lower-casing, stemming, removal of punctuation, digits and stopwords. Lemmatization was not adopted since it considerably decreased the accuracy of the model.
- Topic Features: 20-topic distribution for each article was considered.
- Text Feature Extraction: Dict and Tfidf vectorizer were explored to tranform text. Dict vectorizer was used for its combined ability to perform one-hot encoding and label encoding. Tfidf vectorizer was explored to manage outliers by down weighing words that occurred frequently in the article as well as in the corpus.

The results of each feature set in conjunction with the model will be discussed in the following sections.

2.3 Data Modelling Techniques

To detect climate change misinformation, the following 3 modelling approaches were considered.

1. Outlier Detection
2. Binary Classification
3. Topic Modelling with Binary Classification

2.3.1 Outlier Detection

Initially, the training dataset consisted of only 1,168 examples of climate change misinformation. Therefore, outlier detection algorithms such as One Class SVM and Isolation Forest were investigated. Dict vectorizer was used to transform this dataset. The average precision, recall and F-score for positive class on development set are shown in Table 1.

Although One Class SVM performed slightly better than Isolation Forest, its precision and recall scores were poor on the development set. On an

Classifier	Precision	Recall	F-score
One Class SVM	0.55	0.98	0.70
Isolation Forest	0.51	0.86	0.64

Table 1: Precision, Recall and F-score for outlier detection models.

average only 55% of the time it would identify misinformation. Since the development set consisted of an equal distribution of classes, the model was not robust enough to identify misinformation. One Class SVM resulted in an F-score of 0.46 on the test set, which is considerably low. Hence binary classification was explored.

2.3.2 Binary Classification

The training set was extended with news articles that were collected as part of the data collection phase. Grid search was used to identify suitable parameters that would enable the model to generalize on test data. Of the binary classifiers, Logistic Regression, SVC and Random Forest performed fairly well. It was observed that their performance varied significantly based on the type of vectorizer used. The performance of each model based on feature set and vectorizer are compared against that of the others in terms of F-score, precision and recall. The training dataset utilized for binary classification was transformed using Dict vectorizer and is described in Table 2.

Dataset	Label	Features	No. of instances
Misinformation	Positive class	Preprocess text	1,168
News-articles	Negative class	Preprocess text	1,549

Table 2: Training dataset for binary classification.

Classifier	Precision	Recall	F-score
SVC	0.60	0.98	0.75
Random Forest	0.60	0.96	0.79
Logistic Regression	0.68	0.96	0.79

Table 3: Precision, Recall, F-score for positive class.

The average precision, recall and F-score on development set are shown in Table 3. It was evident that the F-score improved after adding news dataset. However, on test set, SVC yielded an F-score of only 0.47. This indicated that the models did not have sufficient training examples to classify articles that were not misinformation. Therefore, the confusion matrix was inspected to identify articles that were classified as false positives.

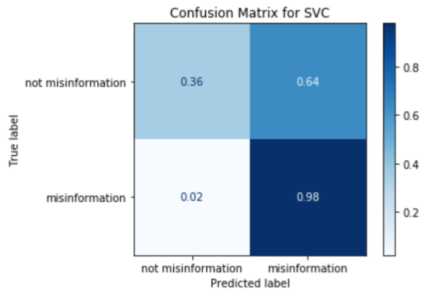

Figure 1: Confusion matrix for SVC

From Figure 1, it was observed that 64% of the articles were classified as false positives. These false positives contained facts of climate change. Furthermore, the news dataset contained few observations related to these facts. This confirmed the assumption that the model did not have enough training examples to identify true climate change information. Therefore, the training dataset was extended to include more examples of climate change from reliable sources. Dict vectorizer was used to transform text.

Dataset	Label	Features	No. of instances
Misinformation	Positive class	Preprocess text	1,168
News-articles + climate change facts	Negative class	Preprocess text	1,139+497

Table 4: Training dataset with Facts for Binary Classification.

The average precision, recall and F-score on development set are shown in Table 5. It was evident that after extending the news dataset with climate change facts, F-score of the classifiers significantly improved. The models were now capable of classifying true and fake climate change news. This was confirmed with the confusion matrix for SVC.

Classifier	Precision	Recall	F-score
SVC	0.86	0.86	0.86
Random Forest	0.81	0.80	0.80
Logistic Regression	0.89	0.88	0.89

Table 5: Precision, Recall, F-score for positive class.

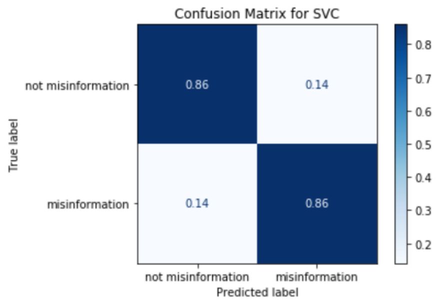

Figure 2: Confusion matrix for SVC

From Figure 2, it was seen that the model was able to identify climate change facts and only 14% were classified as false positives. However, this improvement was a trade-off with increase in false negatives.

To further improve the precision of the model, the dataset was transformed using Tfidf vectorizer. The average precision, recall and F-score on development set using Tfidf vectorizer are shown in Table 6.

Classifier	Precision	Recall	F-score
SVC	0.92	0.88	0.90
Random Forest	0.85	0.80	0.82
Logistic Regression	0.94	0.88	0.91

Table 6: Precision, Recall, F-score for positive class.

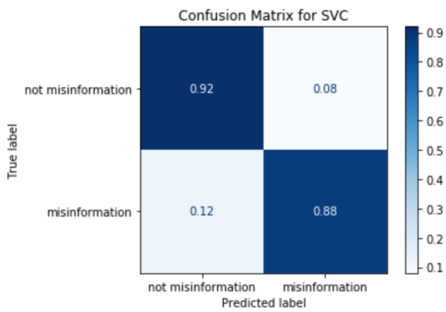

Figure 3: Confusion matrix for SVC

From Figure 3, it is evident that Tfidf vectorization significantly improved the precision of the classifiers. Only 8% of the articles were classified as false positives, and the false negatives were reduced by 2%. This indicated that words such as 'climate change', 'global warming' although important, occurred more frequently in both classes, therefore their weights were reduced and words such as 'fear', 'crisis', 'doom' were given higher weights to identify climate change misinformation. Therefore, the vectorizer was useful in enabling the model to classify true and false climate change news. (Havrlant and Kreinovich, 2017)

Adasyn sampler was used to tackle class imbalance, since recall did not improve, it was dropped.

Other Tested Methodologies with binary classifiers on development set:

1. Standardized numerical features combined with tfidf text features resulted in a slightly lower F-score of 0.89. The numerical features did not significantly improve F-score.

2. Feedforward ANNs with word embeddings, resulted in an accuracy of 88%. However, it resulted in overfitting and performed poorly on test data. The other drawback was that it was resource and time consuming.

2.3.3 Topic Modelling

Topic features seemed more appropriate to classify misinformation since it clusters similar articles together. Genism library was used to pre-process text, Latent Dirichlet Allocation(LDA) was used to model topics from a distribution of 20 topics. These distributions were used as additional features for binary classification. (Ghenai and Mejova, 2017) However, no significant improvement was observed.

From Table 7, it was evident that the topic features did not improve the accuracy of the model. This can be attributed to the fact that LDA considers bag-of-words rather than the context, the distribution seemed skewed towards most frequently occurring words. This can be confirmed from the topic visualization from '*pyLDAvis*'. Words such as 'electricity', 'carbon', 'climate', 'temperature'

Classifier	Precision	Recall	F-score
SVC	0.84	0.76	0.80
Stacking Classifier	0.84	0.64	0.72
Logistic Regression	0.81	0.80	0.80

Table 7: Precision, Recall, F-score for positive class using topic features on development set.

were extensively dominated in both classes. Hence, frequency of words were preferred over context. Therefore, LDA topic distribution did not seem to improve the F-score of the classifiers.

2.4 Evaluation and Results

Based on performance, binary classification was chosen as the most suitable technique. This section describes the error analysis conducted to select a suitable model. The models were evaluated based on ROC curve, precision- recall curve and ability to generalize on test data.

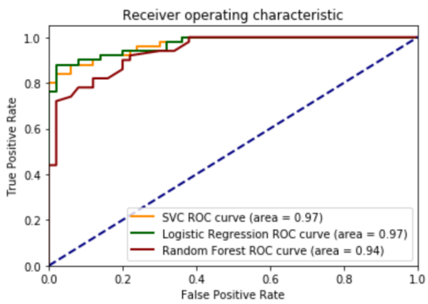

Figure 4: ROC and AUC

From Figure 4, it can be seen that SVC and Logistic Regression performed better than Random Forest, since there area under the curve is greater than that of Random Forest. However, the ROC curves were affected by class imbalance, hence, precision-recall curves seemed more reliable. (Guo et al., 2008)

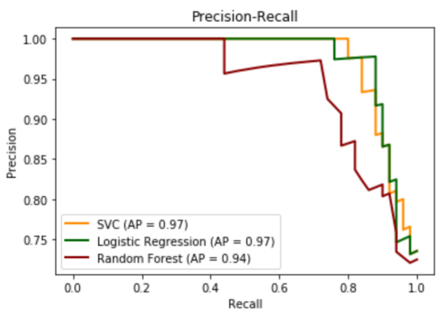

Figure 5: Precision-Recall Curve

From Figure 5, it can be confirmed that SVC and Logistic Regression preformed the best. To select the classifier that generalized better, 10-fold validation was carried out for both models and their precision-recall curves were plotted.

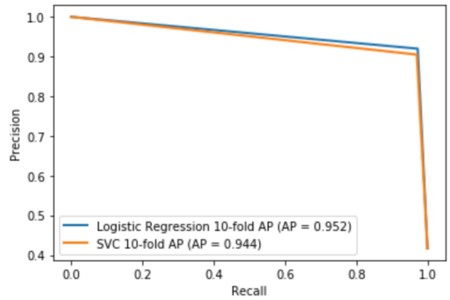

Figure 6: Precision-Recall Curve for Logistic Regression and SVC

From Figure 6, it can be inferred the average precision(AP) for both models differ by a small fraction. Therefore, both the models were trained using training as well as development set. On test set, Logistic Regression yielded and F-score of 0.73 and SVC yielded an F-score of 0.76. Therefore, SVC was chosen as the final model.

In the final evaluation, SVC yielded an F-score of 0.75, as shown in Table 8. The precision has reduced by 0.02 compared to ongoing evaluation. This indicated that the model required more climate change misinformation. Overall, the difference in F-score and precision are quite small. The performance of SVC was as expected and generalized farily well on unseen data.

Classifier	Precision	Recall	F-score
SVC-Ongoing	0.64	0.94	0.76
SVC-Final	0.62	0.94	0.75

Table 8: Precision, Recall, F-score for positive class on test evaluation.

3 Future Work & Conclusion

Future improvements to the project involves scrapping more climate change misinformation and employ techniques to capture the context of the articles, either by using BERT to transform text or extract keywords relevant to each class and train these word embeddings on a CNN classifier. This manner seems more efficient to detect climate change misinformation.

References

https://www.bbc.com/news.

https://edition.cnn.com/.

https://www.theguardian.com/au.

https://www.abc.net.au.

https://www.nasa.gov.

https://www.nationalgeographic.com.au.

https://www.carbonbrief.org.

Amira Ghenai and Yelena Mejova. 2017. Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter. *CoRR*, abs/1707.03778.

X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. 2008. On the class imbalance problem. In *2008 Fourth International Conference on Natural Computation*, volume 4, pages 192–201.

Lukás Havrlant and Vladik Kreinovich. 2017. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *Int. J. General Systems*, 46(1):27–36.