Climate Change Misinformation Detection System

Wenkang Zhu **Student Number: 958668**

Abstract

The Climate Change misinformation detection system development competition was launched to identify misinformation from the various raw test. A data set containing both misinformation and non-misinformation texts and labels are provided by the organizer. In this report, several designs of the detection system are presented. The performance and defects of these systems are analyzed and discussed.

1 introduction

Many misinformation news and texts regarding the environment and climate change have appeared on the internet and websites in the past years. Climate change misinformation is hard for the reader to do fact-checking before they read them. A large number of climate change misinformation texts are provided by the organizer so that climate change misinformation detection system can be built based on those data. Several systems are built in this report. Methods such as Bag of words (BOW) and Bidirectional Encoder Representation from Transformers (BERT) are used to extract features from text data. One class support vector machine is used to identify non-misinformation and misinformation texts. Binary classification method such as BERT classification is implemented.

2 Background

This section describes the methods and machine learning models used to develop the misinformation detection system.

2.1 Feature extraction methods

Bag of Word is one of the most popular and easy models used in natural language processing and information retrieval to extract features from text data. Every distinct word is used as a feature and the occurrences of the words are the feature data for each instance/document(Harris, 1954).

Truncated SVD is a popular method to reduce dimension after feature extraction in natural language processing and it is similar to PCA. Unlike PCA, Truncated SVD does not center data before computing(Xu, 1998).

BERT is a new and powerful language representation model. It has a pre-trained deep bidirectional model from a large unlabeled text dataset that can represent the contextual meaning of words and documents. This model can be fine-tuned with an additional layer for various language tasks.(Devlin et al., 2018).

2.2 One class Support Vector Machine

Since we only have texts with label 1, which is misinformation texts, in our training data. One Class SVM method can perfectly fit this type of task. It is an unsupervised kernel method that supports high-dimensional distribution. Only one class of data is required. It is similar to the SVM method in supervised learning. This method uses training data to generates a binary function to identify whether the novelty's probability density region is in that region of the training data. Novelty is an outlier if it is not.(Schölkopf et al., 2001).

3 Preprocessing and some analysis

This section how the preprocessing is done and some basic data analysis.

3.1 Preprocessing

The regular expression is used to remove most of the punctuations and website addresses. Question marks and exclamation marks are kept since those environmental misinformation texts tend to use that two punctuation to present angry and shock. The numbers in the text are removed. All entity names, such as organization names and personal names,

are also removed since they appear frequently in both classes and bring some noises. Words' length less than 2 is deleted. All texts/documents are tokenized and lemmatized. Some words are processed elaborately, such as replace ”it's” to ”it is” and ”I've” to ”I have”. There is some encoding error lead to string such as ”
xa0änd ”
u200b” are also removed. Various encoding types are tied to avoid this situation when open the JSON file. However, Those error string seems to come from the data set itself. In the end, words are represented in the form of Bag of Words.

3.2 Some Simple Analysis

After preprocessing all the text, a word cloud is plotted according to all the words in the training data set. Some words such as carbon dioxide, fossil fuel, and global warming are very frequent in misinformation texts. All those texts are related to climate change and the environment. A word cloud regarding non-misinformation texts in development data is also plotted:

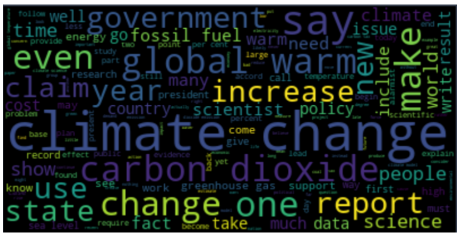


Figure 1: word cloud of misinformation text

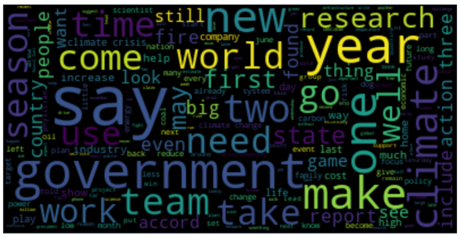


Figure 2: word cloud of non-misinformation text

Fewer words regarding climate change and environment appear in this plot. Contents and topics of non-misinformation texts tend to be more varied. There are obvious words and tones in the misinformation text. Most Frequent words plot also indicates that misinformation has some pattern in the choices of words.

The development data set to share some common most frequent words with training data since there are both misinformation and non-misinformation text in it.

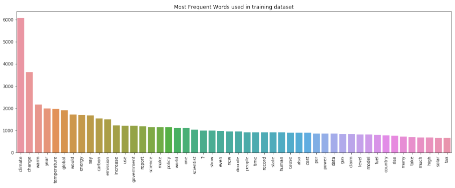


Figure 3: Most Frequent words in Training data set

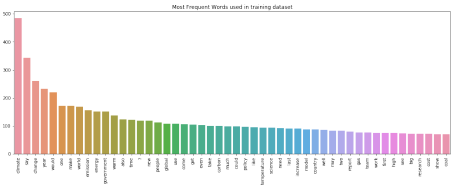


Figure 4: Most Frequent words in development data set

However, test data show a different pattern with the other two data sets. Most frequent words in the test data set appear to be more random. Seems it is hard to identify what the whole test set is related to.

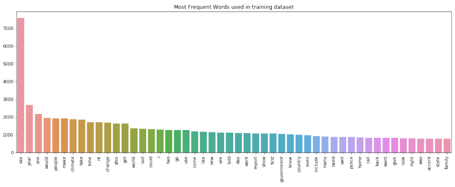


Figure 5: Most Frequent words in test data set

The test data set show a different word pattern with training and development data. Systems are expected to have worse performance on test data.

4 systems for Climate Change Misinformation Detection

This section describes several models that are developed for Climate Change Misinformation Detection System and their performance on development data.

4.1 One Class SVM

Baseline One Class SVM: Climate Change Misinformation Detection is treated as one class classification task initially since we only have one class data for our training data. Bag of words method is used for text feature extraction in this baseline system. Extracted features are trained based on One class SVM (herein OCSVM) and tuned according to performance on development data. In this case, non-misinformation text(label 0) is treated as outlier, and misinformation (label 1) is treated as inliers.

Truncated SVD + OCSVM: Bag of words leads to a large feature matrix that may cause the curse of dimensionality. Over 19000 features are extracted by the BOW method from our training data. To avoid this problem, Truncated singular value decomposition (SVD) is used to reduce the dimension of the data. Feature dimensions of our data is reduced to 300 by cross-validation. Reduced training data are trained in one class SVM model and tuned to get the best performance on development data.

BERT+OCSVM: Bag of words is a very popular feature extraction technique. However, it cannot represent the meaning of each sentence and word based on its context. To better represent the contextual meaning of each document, BERT is used to do a word-embedding job. All text data are preprocessed into BERT inputting format (add special tokens, padding and convert to ids, etc.,). The maximum length of each preprocessed instance is padded or truncated to 512 since this is the maximum length that BERT can take. Processed data are passed into the pre-trained model ”bert-base-uncased”. Extracted features with 768 dimensions are obtained since this is the dimension of the last hidden state of the pre-trained BERT model. Extracted features from BERT are reduced in dimension and trained in one class SVM model.

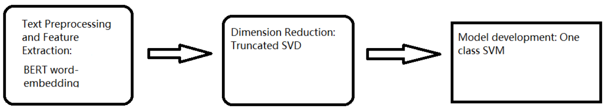

Text Preprocessing and Feature Extraction:
BERT word-embedding
Dimension Reduction: Truncated SVD
Model development: One class SVM

Figure 6: A simple flow chart of the system: BERT+OCSVM

4.2 BERT Classification

Unlike the system stated above, this task is treated as a binary classification problem rather than one class classification and anomaly detection. Extra non-misinformation data are added to training data. To avoid data leakage problem, extra data are past CNN news downloaded from website(See et al., 2017), which means their source are different from raw training data provided. For the model architecture, to prevent over-fitting, only one additional dropout layer and one classifier layer are added to the pre-trained BERT model. The whole BERT model will output a logit result for each instance.

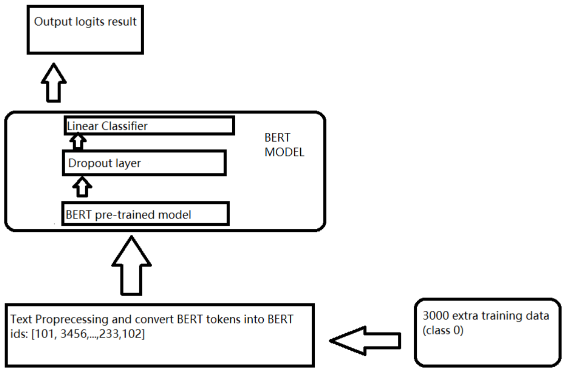
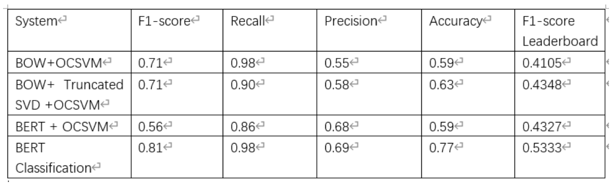

Output logits result
Linear Classifier
Dropout layer
BERT pre-trained model
BERT MODEL
Text Proprecessing and convert BERT tokens into BERT ids: [101, 3456,...,233,102]
3000 extra training data (class 0)

Figure 7: A simple flow chart of the system: BERT Classification

5 Result

The best result of each system implemented in this report is shown. These results are the performance on climate change misinformation class (label 1) in the development data provided. Parameters of each system are tuning by the grid search method. the baseline system (BOW + OCSVM) have not bad performance on development data provided and have 0.71 on F1-score. Compare to the baseline, the second system with dimension reduction does have a little improvement on accuracy and F1-score on Coda-lab leader board. However, it is not a very obvious improvement. Word-embedding by BERT does not have a good performance as expected. The result is even worse than BOW for feature extraction. This situation may cause by inappropriate preprocessing and maximum input limit of the pre-trained model. Most of our text data are longer than that after tokenization. Some information in the text is truncated as well. The other reason may be that pre-trained BERT does not have a good performance on some word-embedding tasks. With more training data, BERT Classification system provided the best performance on development data. F1-score on development data reached 0.81 which is the best performance till now. Also result on Coda-lab leader board final evaluation reached 0.52 for



System	F1-score	Recall	Precision	Accuracy	F1-score Leaderboard
BOW+OCSVM	0.71	0.98	0.55	0.59	0.4105
BOW+ Truncated SVD +OCSVM	0.71	0.90	0.58	0.63	0.4348
BERT + OCSVM	0.56	0.86	0.68	0.59	0.4327
BERT Classification	0.81	0.98	0.69	0.77	0.5333

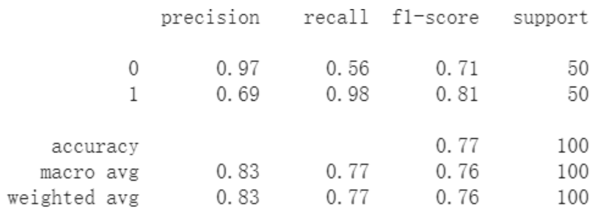
Figure 8: Result of different systems

F1-score, which is the highest one among all my submissions.

This result still needs to improve compared to the other teams. Some aspects still need to improve such as model architecture and scrape more high-quality text data. BERT is expected to have a better performance on shorter sentences(Yang et al., 2019). However, most of the method are complicated and hard to implement. Improvement on performance is not guaranteed as well. To solve this problem, text summarizing should be applied to the raw data so that they can be built on summarizing text. Due to the time limit, this thought is not implemented.

6 Error Analysis

Result of the best system in this report, which is BERT classification, is stated below.



```
              precision    recall  f1-score   support

           0       0.97      0.56      0.71        50
           1       0.69      0.98      0.81        50

    accuracy                           0.77       100
   macro avg       0.83      0.77      0.76       100
weighted avg       0.83      0.77      0.76       100
```

Figure 9: Result of BERT classification

According to the result on development data, the classifier has a high recall (0.98) on identifying misinformation text, which means only 1/50 misinformation (class 0) text is identified as non-misinformation (class 0) text by the classifier. However, Recall of class 0 and precision of class 1 is not high enough. This means 22/50 non-misinformation texts are classified as misinformation text by the classifier. Many of the misclassified non-misinformation texts are related to climate change and the environment. Below is an example of a non-misinformation text that is not correctly identified.

> ’The government must abandon its fossil fuel power projects. If not, we’ll sue. No longer should our survival be an afterthought. If we are to withstand the climate crisis, every decision should begin with the question of what the planet can endure. This means that any discussion about new infrastructure should begin with ecological constraints. The figures are stark. A paper published in Nature last year showed that existing energy infrastructure, if it is allowed to run to the end of its natural life, will produce around 660 gigatonnes of CO2.

This example is related to climate change. Seems like the classifier can only identify whether the text is related to climate change or environment. Fail to distinguish between misinformation and non-misinformation for all texts related to climate change. Below is a misinformation text example.

> Climate Strike Kids Cool on Real Action A popular rebuttal to the Klimate Kiddies is that they should walk the walk, give up their power-gobbling devices and stop riding mummy and daddy’s gas-guzzling 4WD to school. Another way these shrieking ninnies could put their money where their mouths are is to give up air-conditioning. That’s not about to happen. In fact, quite the opposite. More than 1300 schools have put their hands up for airconditioning under the Cooler Classrooms program, but almost a year after the first round of recipients was announced, just 27 have so far received it.

It is even a little hard for a human to identify whether that environment-related information quoted above is misinformation. However, we still can tell there is not a very obvious difference from the tone of the text. The model is expected to have a great performance if this difference can be captured.

7 Conclusion

In this report, several systems are presented for climate change misinformation detection, which includes one-class classification method and binary classification method with extra data. The system mainly contains 3 part, text data pre-processing, feature extraction, and model development. One class SVM with different feature extraction methods and BERT classification is implemented for these tasks. However, these systems do not provide very good results on Coda-lab leader board. It is expected to get a better result if BERT input limit problem can be solved properly with a more elaborate parameter tuning and more training data.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Peiliang Xu. 1998. Truncated svd methods for discrete linear ill-posed problems. *Geophysical Journal International*, 135(2):505–514.

Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*.