COMP90042 Report for Project 1

Ge Yang 925156

Abstraction

Misinformation in climate change is prevalent in mass media. This paper aims to explore how different skills can be implemented on a classification problem regarding this issue. Given data sets and external data sets are introduced along with pre-processing and model selection. Different classifiers are tested. Results based on different configurations of classifiers and models are discussed. Configuration of logistic regression and bigram model gives best performance in this study. Error analysis shows different behaviours of different models and classifiers.

1. Introduction

Climate change counter movement (CCCM) organizations propel crafting and spreading misinformation concerning climate change for various reasons [1].

Apart from intentional manipulation, dramatic story which often involves misinformation draws more exposure, hence more motivated. It is a common practice for the media to interpret climate issues unprofessionally. Penalty on commensurate false alarm is literally none [2].

The purpose of this report is to help identifying texts with misinformation. A binary classification approach is adopted.

One class classification approach often triumphs when a particular class has much more samples than the other. In this study, large number of texts with negative labels are available, so there is no discrimination towards one class or imbalance problem in dataset. Hence no necessity for one class classification.

2. Dataset

Training data including 1168 texts with uniform positive labels while 100 texts in development data with mixed labels. Positive label indicates misinformation exists in the texts. The intuitive may be that texts with negative labels are easier to find and scraped.

In this study, external texts with negative labels are sourced from newspapers such as BBC, CNN, Fox, Washington Post and etc. A crawler called **newspaper 3K** is used. The scraping procedure ensures each source contributes equally to the number of samples in the final external dataset.

First dataset is a large mixture of climate related and unrelated texts. It is achieved by scraping the local content and every link on the current page. Average number of words in each text is 479. We call this dataset number 1.

Second dataset is specifically climate related texts. However, this dataset is small with only 100 texts in total with an average 1263 words for each text. We call this dataset number 2.

3. Pre-processing and Models

3. 1. Pre-processing

Pre-processing includes several steps. Firstly, text and label of each data set are placed in two separate lists for training purpose. Then each text set is tokenized, stripped out of stop words. Whole text is then modified by regular expression created to reserve meaningful sequence. Lemmatization is also added for bag of words model specifically with pos-tag as reference.

3. 1. Models

First way is to tokenize every word in text and accumulate them into a **bag of words**. Each piece of text has a unique bag of words which could then be vectorized. This skill helps to identify the variety and abundance of words within the text. Second way is to use a **n-gram** paradigm. Vectorization is based on the n-gram dictionary. In this study, bigram and trigram models are tested. Lastly, a **TF-IDF** model is used.

3. Classifiers

3. 1. Logistic Regression

This is the most preliminary but may also be the most effective classifier for this problem. Logistic regression produces binary prediction on given input. Training of logistic regression classifier is swift and computationally cheap [3]. Different values for regularization parameter C are tested on development dataset.

3. 2. SVM

Support vector machine (SVM) aims to perform binary and multi classification. Performance normally depends on kernel options and other parameters. This study trials linear, polynomial, gaussian RBF kernels. Parameter tunned is regularization parameter C.

3. 3. LSTM

LSTM is a mutant of GRU (Gated Recurrent Unit) with three **gates** for forget, update and output function to maintain "memory" in a sequence. It takes in current input (word) and the output gate coefficient and stage value before softmax from previous cell to study the information flow.

In this study, a simple structure of the network is used including one embedding layer plus one LSTM layer and one output layer with dimension one to output classification results with sigmoid activation. Adam optimizer is chosen to reduce the plateau impact.

However, LSTM can have poor performance with long input sequence and short outputs such as classification problem [4].

4. Results and Discussions

4. 1. Models and Classifier

Following chart shows performance of different features under different classifiers. Detail configurations are stated.

F1 (%) Dev	LGR	SVM
Bag of Words	82%	81%
Bigram model	90%	85%
Trigram model	90%	85%
TFIDF model	82%	88%

Table 1: F1 score for different features and classifiers on development dataset. C=10 in SVM and Bag of words is not lemmatized by pos-tag. Only unigram is in the bag.

It could be seen that bigram model with logistic regression configuration outperforms other configurations. Trigram model does not increase performance compared to bigram. Bag of words model is the one with worst performance. This is understandable since it only harbors unigram in this implementation. Thus, less information within sequence is encoded and analyzed. TFIDF model surprisingly maintains a good performance.

For LSTM, the results are shown in table 2. Although it seems that the performance generally resembles the level of the traditional methods. However, the F1 score tested on CODALAB with partial test dataset plunged significantly compared to other methods.

LSTM (%)	F1	Precision	Recall
	82	71	98

Table 2: F1 score for LSTM with bigram model on development dataset.

LSTM (%)	F1	Precision	Recall
	47	30	100
SVM (%)	F1	Precision	Recall
	67	52	94

Table 3: F1 score for LSTM and SVM with bigram model on ongoing test dataset.

It shows that this LSTM configuration **overfits** severely on the development dataset. This verified that LSTM is not good at classification problem. As a matter of fact, RNN with **multi to one** configuration is more appropriate for classification problem theoretically since there is only one output after an epoch that embodies whole information flow along the propagation.

4. 2. External Datasets Trial

Table 4 shows the results comparison between big dataset No.1 (mentioned in datasets) as sole external datasets and two datasets combined as external datasets.

External Dataset	No.1	No.1 & No.2
F1 (%)	79%	84%

Table 4. F1 score of Dataset No.1 as sole external dataset vs. No.1 and No.2 datasets combined as external dataset.

With the small but concentrate dataset highly related to climate change, the f1 score increased by 5% in development dataset. This increase extends to ongoing test dataset with a magnitude of 3%. So, it is safe to say **highly climate related** content can be beneficial to the training. This is based on the fact that originally large external dataset has mixed content.

4. 3. Performance in final test

As table 5 shown, final rank is 82 among 282 submissions. F1 score on final test decreased by 22% compared to development dataset. The gap is stable throughout all methods tested previously.

LGR & Bigram (%)	F1	Precision	Recall
	68	53	94

Table 5: Best performance in final test using logistic regression and bigram model.

5. Error Analysis

Table 6 shows error types distribution for development dataset. When logistic regression classifier is used throughout all models, TFIDF model produce higher ratio of **'1-0'** errors, meaning mislabeling positive text as negative one.

Error	TFIDF	Bigram	Bag of Words
1 to 0	11	3	5
0 to 1	8	8	14

Table 6: Error types when logistic regression used throughout three models for development dataset. '1-0' means correct label is 1 and false prediction is 0, vice versa for '0-1'.

Bigram on the other hand has far less '1-0' errors. Bag of words model has an opposite error ratio compared to TFIDF model.

5. Conclusion

In this report, several NLP models are used and tuned. **Logistic regression** combined with **bigram model** performs the best both in development and test dataset. LSTM is found to be not particularly suited for long text classification problem even after tuning. High concentration of climate change related samples has huge influence over the final result. Error analysis shows subtle differences among language models.

A more delicate division should be carried out to address the ratio of climate change related samples within external dataset. Ambition is to collect both pure climate related dataset and unrelated dataset. Mix them with a certain **ratio** and use the ratio as a hyperparameter in tuning. More 'classification-friendly' RNN configuration such as **many inputs to one output** configuration should be tested for this task.

6. References

[1]. Justin Farrell. 2016. Network structure and influence of the climate change counter-movement. *Nature Climate Change*, 6(4):370.

[2]. Riley E Dunlap and Peter J Jacques. 2013. Climate change denial books and conservative think tanks: Exploring the connection. *American Behavioral Scientist*, 57(6):699–731.

[3]. Alexander Genkin, David D Lewis & David Madigan (2007) Large-Scale Bayesian Logistic Regression for Text Categorization, *Technometrics*, 49:3, 291-304

[4]. Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," in *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68-75, 3 2017