Offensive Tweet Classification

June 22, 2019

陳君彥, b04703091 b04703091@ntu.edu.tw
陳柔安, b04701232 b04701232@ntu.edu.tw
蕭法宣, b04705007 b04705007@ntu.edu.tw

Division of Work

陳君彥, b04703091:

- Generation and testing of TFIDF-Vectorized models.
- Generation and testing of domain knowledge based features.
- Creation of written report.

陳柔安, b04701232:

- Generation and testing of bi-LSTM.
- Generation and testing of CNN models.

蕭法宣, b04705007:

- Generation and testing of BERT model.
- Generation and testing of string based features.

1 Introduction

The goal of OffensEval competition on codalab by SemEval 2019 [1] is to tag a series of tweets with regards to their offensive nature. The competition is separated into 3 sub-tasks.

- Sub-task a: Tag a tweet based on whether the tweet is considered ”offensive” or not. Tags include ”OFF” and ”NOT”
- Sub-task b: If a tweet is tagged as offensive in sub-task a, further tag the tweet on whether it is offensive in a targetted entity, or not targetted. Tags include ”TIN” and ”UNT”.
- Sub-task c: If a tweet is tagged as targetted in sub-task b, further tag the tweet on whether it is targetted towards an individual, towards a group, or towards something else. Tags include ”IND”, ”GRP”, and ”OTH”.

1.1 Dataset Analysis

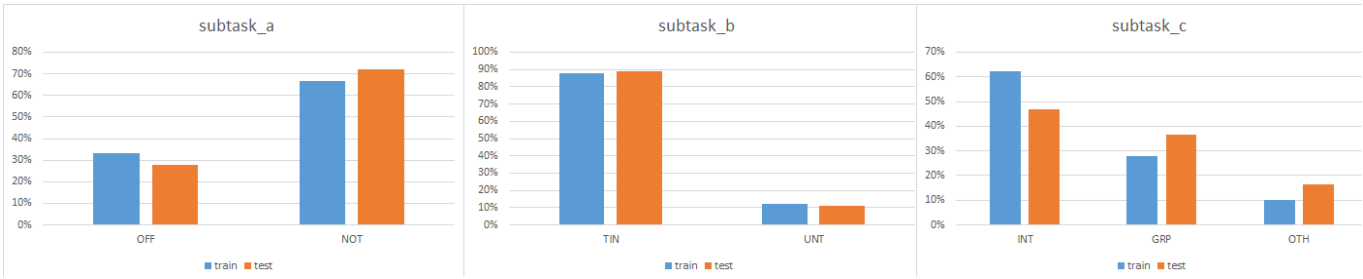

Figure 1: The percentage of each tag in the data set

Our training dataset consists of 13240, 4400, 3876 tweets for sub-task a, b, c respectively, and the testing set has 620, 240, 213 respectively. We suspect the small size in the training data, coupled with the large disparity between the tag distribution as seen in Figure 1, contributes greatly to the difficulty in creating an accurate prediction model.

2 Conclusion

We arrive at the conclusion that the aggregation of different features in our data is an effective method in indentifying and classifying fake news. After discussions with classmates and scouring forums for other people's attempt, we also conclude that this method is also rather efficient.

Our personal machines were all laptop spec'd machines, running the models within an acceptable time frame, with the results being only slightly lower than much more intensive models, such as BERT, which often require much more powerful hardware, and ran upwards of hours when training the models and predicting the results.

References

[1] SemEval. (2019) SemEval - OffensEval: Identifying and Categorizing Offensive Language in Social Media. [Online]. Available: competitions.codalab.org/competitions/20011