Analysis for Paraphrasing tweets to Enhance Tweet Reliable/Unreliable Classification Accuracy

Yu-Jui Chang

Introduction

Since the tweet spreading fast, sometimes, it even propagate faster than news. However, people post the tweets not like journalist post news needed supervisor agree, they can post anything they like. Therefore, it will cause some issues. Such as, fake news or something do not proved truth are the most seeing. Because internet spread fast and people usually not check the sources' reliability, it will mislead people easily if someone want to do so. If we can build a model to distinguish reliable or unreliable tweets, it would help people with reading tweets.

- Using '*COVID19_Tweet Dataset' train our model.*
- Check accuracy before and after extending dataset with paraphrasing.

How to paraphrase

●Use **NLTK** module to paraphrase. First, check the POS for each word.

●Second, check each word synonyms existing or not.

●Third, random choose word synonyms to build new sentence.

●Example :

original: ***The quick fox jumps over the dog***

['The', []]

['quick', ['fast', 'promptly', 'quickly', 'speedy', 'agile', 'ready', 'warm', …]

['fox', ['Charles_James_Fox', 'slyboots', 'George_Fox', 'dodger', 'fox', 'Fox']]

['jumps', ['leap', 'jumping', 'saltation', 'parachuting', 'startle', 'jump', 'start']]

['over', []]

['the', []]

['dog', ['frankfurter', 'wienerwurst', 'hot_dog', 'blackguard', 'hound',…]]

paraphrase: ***The fast slyboots jumping over the hound***

Result

Accuracy	Before paraphrasing (560 sentences)	After paraphrasing, (2520 sentences)
Logistic Regression	81.2%	**83.5%**
SVM	80.7%	**86%**

Experiment

- We have a dataset with 560 tweets including 280 reliable and 280 unreliable tweets.
- Step1:
- Data preprocessing
- Use *TFIDF* to build the *Logistic Regression/ SVM* model through 10 folds cross validation check the preliminary classification accuracy
- Step2:
- Paraphrase 70% dataset sentences (560*0.7)
- Paraphrase each 70% sentence to 5 different sentences. Extend dataset from 560 to 2520 (560*0.7 + 560*0.7*5 + 168(=560*0.3))
- Step3:
- Re-train the new dataset(2520) and check new classification accuracy.

Conclusion and future work

Both models enhance accuracy. However, both methods doesn't enhance really great. The reason with not enhance large accuracy may because of paraphrasing sentences features too similar with original sentences, so the training model cannot get improve on prediction. The model doesn't get any new features.

In the future, we may want to try another paraphrase method. Try to paraphrase the sentence using different structure because our way has same structure and only change the vocabulary. If we can creature a new sentence with different structure, it might get the better classification accuracy.

This project is for STAT-696-Natural Language Processing class, American University. Idea is from Li, C., Su, Y., & Liu, W. (2018, July). Text-to-text generative adversarial networks. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE. I tried to use paraphrase to replace generator part with creating new sentences. Then tested classification accuracy with different classifiers and found which one is best.

1. Li, C., Su, Y., & Liu, W. (2018, July). Text-to-text generative adversarial networks. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
2. Boukouvalas, Z., Mallinson, C., Crothers, E., Japkowicz, N., Piplai, A., Sudip, M., ... & Adalı, T. (2020). Independent Component Analysis for Trustworthy Cyberspace during High Impact Events: An Application to Covid-19. *arXiv preprint arXiv:2006.01284*.