Proposal

We picked a competition named “Real or Not? NLP with Disaster Tweets” that predict which Tweets are about real disasters and which ones are not from Kaggle. Here’s the link:

https://www.kaggle.com/c/nlp-getting-started/overview

Twitter has played an important role in social media life. Nowadays, people like to post a Tweet to share things and many people get information from Tweets, like news, weather. For example, some people may post a Tweet when they see disasters to alarm other people. However, not every Tweet is about real disasters, sometimes people use words like “ablaze” metaphorically to describe the beautiful view. If we can use these real-time Tweets efficiently, it may help people avoid some unnecessary harm.

About the dataset, the train and test data consist of tweets, which may contain disaster information. Since these data contain some symbols, website information, abbreviation and computer cannot read words directly, the dataset needs to be cleaned first.

In our plan, we will use the traditional machine learning algorithms: KNN, K-means, SVM, Decision tree, Logistic regression and Naive Bayes. What’s more, since this problem is a classic NLP binary classification problem, we are going to use the deep learning model, Bert, which developed by Google aiming to solve many NLP problems. Finally, we will combine these methods and use a boosting model to improve our model. For instance, random forest. In the meantime, all of our algorithms are in standard form.

As for software, our bert model will run on Google Colaboratary, which have a free GPU to support the code, since our computer cannot support the computation of deep learning model. Other algorithms will use PyCharm to run. And we will use Bert-as-service to encode our text information, because we need to convert the text to number to support our algorithm. Concrete information about Bert-as-service will provide later.

As for reference materials, we reviewed lectures in the classroom, browsed a large number of relevant information webpages, and read relevant books and literature. Through reading and discussion, we understand the impact of social networks, what is NLP, how to apply python to NLP, and how to analyze and judge the results. This laid a good foundation for us to successfully complete the follow-up research. Related references are listed at the end of the article.

Except error rate and F1-score, there are two methods applied to evaluate and compare the performance of different classifiers:

Precision-Recall (PR) Curve and Receiver Operating Characteristic (ROC) Curve. In both curves, the area under the curve (AUC) will be used as a summary of the model skill. The greater the area means the better performance. Using ROC curve, different models can be compared directly in general or for different thresholds.

Finally, there’s our schedule:

4.5-6: Group meeting [1]- Assigning the tasks. Finish the proposal.

4.7-12: Individual work.

4.13: Group meeting [2]- Integration and discussion.

4.14-15 Revise and refine.

Reference

1. In disasters, Twitter influencers get out-tweeted. (2019, February 13). Retrieved from https://www.sciencedaily.com/releases/2019/02/190213142654.htm
2. Umihara, J., & Nishikitani, M. (2013). Emergent Use of Twitter in the 2011 Tohoku Earthquake. *Prehospital and Disaster Medicine*, *28*(5), 434–440. doi: 10.1017/s1049023x13008704
3. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing Text with the Natural Language Toolkit.* Beijing: OReilly.
4. Goyal, P., Pandey, S., & Jain, K. (2018). *Deep learning for natural language processing: creating neural networks with Python*. Berkeley, CA: Apress. doi: 10.1007/978-1-4842-3685-7
5. Real or Not? NLP with Disaster Tweets. (n.d.). Retrieved from https://www.kaggle.com/c/nlp-getting-started
6. Brownlee, J. (2019, December 18). How to Use ROC Curves and Precision-Recall Curves for Classification in Python. Retrieved from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/
7. ZHOU, Z. H. I.-H. U. A. (2020). *Machine Learning*. S.l.: SPRINGER VERLAG, SINGAPOR.
8. Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding