Tweet Sentiment Classification Proposal

Jitian Zhao jzhao326@wisc.edu
Chen Xing cxing6@wisc.edu
Lize Du ldu43@wisc.edu

1. Introduction

This project will mainly talk about how to do sentiment analysis using deep learning methods. Our goal in the end is to train a neural network to tell us whether an input sentences attitude is negative, neutral or positive. To achieve this goal, we will simply split our project to 4 parts:

- Data cleaning.
- Train our neural network.
- Avoid over-fitting using regularization methods.
- Compare our model with models built with machine learning methods.

Considering that we haven't learned the main methods we are going to use, data cleaning is the part we discussed most for now. We brainstormed some critical case that may influence our prediction outcome a lot, and for each possible problem, we come up with some solutions to deal with it.

- **Some words are just different tense for the same meaning.**
  For example like:liked, liking, interest:interested, interesting, interests.
  NLTK package has PorterStemmer() to do stem extract and WordNetLemmatizer() to transfer the verb to its normal form.
- **The same word in different sentences can have different meanings.**
  For example: I like this course and I don't like this course. For the same word like,the two sentence have two opposite meaning.
  Add _neg label between not/never and the first punctuation after it, using NLTK sentiment tools.
- **Some words have a huge frequency, but they are actually useless.**
  For example, the, is, and some words like these can not help us to predict the attitude but they have a large word frequency, we just want to keep informative words! We can see this from the word cloud figure in figure 1. Words that appear frequently might not contain much information.
  We decide to use information gain to select some top words and use Tf-Idf method to create the word vector.

Figure 1. word cloud for the text

- **How to tokenize sentence?**
  We want to use regular expression and the tokenize method in NLTK, because if we just match words, we may miss some words like built-in or U.S.A., i.e. the words which are not just composed of character but some symbols.

As for training the neural network. We consider using RNN, CNN and LSTM. We will introduce these methods after they are covered in class. Since over-fitting may do harm to our predict mechanism, we will improve our model by adding regularization to it. And to show the advantage of deep learning method, we will build models using simple machine learning method such as naive bayes and compare the accuracy.

There are some work on the internet that are related to this topic. For example, Zhang, Wang and Liu conducted a survey in 2018.[3] This essay briefly described the deep learning methods that are commonly used in sentiment analysis and listed some existed work related to it. After reading this, we learned how the procedures are designed in sentiment analysis.

2. Motivation

Sentiment analysis or opinion mining is the computational study of peoples opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.[2] Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing (NLP).Nowadays, if one wants to buy a consumer product, one is no longer limited to asking ones friends and family for opinions because there are many user reviews and discussions about the product in public forums on the Web. However, finding and monitoring opinion sites on the Web and distilling the information contained in them remains a formidable task because of the proliferation of diverse sites.[3]
This is also true for Twitter, a popular micro-blogging service where users create status messages (called tweets). These tweets sometimes express opinions about different topics.[1]There are many information contained in the tweets and we want to construct a model that can extract these information and to detect whether the tweet is positive or negative.

3. Evaluation

Since the target variable in our model is whether the tweet is positive or not, it is actually binary classification problem. Since the number of sample is quite larger in our problem, we mainly focus on test accuracy to estimate the performance of the model.
Meanwhile, we also look at the gap between training accuracy and test accuracy so that our model won't have a large over-fitting problem. We also use the confusion matrix to further illustrate the performance of our model.

4. Resources

Datasets

The dataset we use is Sentiment 140 which was created by Alec Go,Richa Bhayani,Lei Huang who were Computer Science graduate student from Stanford university.
The dataset contains lots of features but we only use the polarity of the tweet and the text of the tweet in our model. In particular, the polarity of the tweet has 3 different kind of labels:negative, neural and positive.

Tools

In our project, we implement some machine learning methods as baseline and some deep learning method.
In particular, for machine learning method, we plan to use sklearn python package and for deep learning method, we plan to use pytorch.

5. Contributions

Data cleaning: Lize Du
Deep learning model constructing: Jitian Zhao, Lize Du,Chen Xing
Machine learning model constructing: Jitian Zhao, Lize Du,Chen Xing
Report writing: Jitian Zhao, Lize Du,Chen Xing

References

[1] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[2] B. Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[3] L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.