Weekly Plan

1. Set report format
2. Finish reading last two papers
3. Update the literature review of the three papers (according to the literature review format)
4. Set Colab
5. Find a suitable dataset and run vader

1. Report format

I used the template offered by the course on wattle and uploaded to overleaf online editor. The following report writing should be done on overleaf

2 & 3. Literature review

The draft has been uploaded to gitlab

Lexicon:

The role of lexicon in natural language processing

This article mainly introduced the history of the development of machine reading texts and is divided into different parts: the early work, taxonomies implicit in dictionary definitions, extracting information from LDOCE, extracting information from LDOCE, building a lexical database, database approaches to the machine-tractable dictionary, lexical disambiguation and creating bilingual lexicons for NLP projects, and the relation of MRDs and corpora. According to the article, lexical is part of NLP system that contains semantic and grammatical information about individual words or word strings, and can provide a structure which not only change the dictionaries into machine readable ones, but also can pick the out the right meaning according to the text. The article briefly described how the traditional dictionary explain words, like using a genus term and then exclude other meanings, and also used Longman Dictionary of Contemporary English to realize how to point out the right meaning, like going through the list and compare the meaning with the rest of the text, and pick the one with the highest overlap rate. Comparing with the corpus, lexicon is more like dictionary: a list of text, and corpus basically means a body of text. One of the main fields of lexicon application is sentiment analysis.

Lexicon Generation for Emotion Detection from Text

There are three different techniques commonly used to build a lexicon. First of all, the traditional manual lexicon, which can be constructed by human hand, all the words contained are graded by different people. Second, the ontology based lexicon, which is generated according to the external relations of known seed words and their related synonyms, antonyms and supersemes, spreads objects on the graph. Thirdly, corpus based lexicon, mainly involves conditional probability and pointwise interoperability. However, this technique is more likely to overemphasize the relationship between terms and classes. At a simple statistical level, if a word occasionally appears too much in a class, although the word usually has no trend of any class, it may mislead its weight in the vocabulary and make it tend to that class.

Classifiers

The construction of classifier is based on the algorithm supporting classifier. Different types

of classifiers may have their own characteristics, they have their own target regions and data sets. Therefore, it is necessary to test the algorithm to determine whether the selected classifier is suitable for the data to be processed.

Lexicon-based Classifiers

The classifier based on lexicon can be represented as a dictionary like table, in which the weight of each word is stored. Weights are used to measure, for example, how a word affects a sentence, or in other words, how a word in a sequence affects the trend of the category to which the entity belongs.

Towards Explainable Text Classification by Jointly Learning Lexicon

Lexicon based lexicon can be described as an open source dictionary, people can view detailed information at any time. In addition, because its foundation is built manually, it can be easily modified and adjusted when needed. On the other hand, it is a hard work to build a complete vocabulary, because there are more than 15000 English words in use, and many of them have different meanings, which may take a lot of time. Although a dictionary has been successfully constructed, it takes time and effort to test and evaluate it on different types of data, and its accuracy cannot be guaranteed at the same time.

4. Set Colab

I applied for account of google colaboratory, which may be useful for future coding. The demo, however, can run currently on my machine.

5.Vader Demo

Shixuan provided three datasets: an IMDB comment dataset, an Amazon comment dataset and a Yelp comment one. I also used the poem *The Raven* as a dataset. Vader works well on the comments, but did not do as well as that on poem. I guess it may be a shortage for lexicon-based sentiment analysis as reading poem requires far beyond analyzing vocabulary. While for comments the emotions are more direct and intense.

The demo, including codes, datasets and running result is uploaded to gitlab.