Natural Language Processing, Project Report

UID: 1690550

March 16, 2017

1 Introduction

Sentiment classification is an ongoing area of research in data mining and natural language processing areas. It has a wide range of applications, varying from opinion gauging to marketing. Its potential is constantly growing together with the development of related areas, including machine learning and neural networks.

The notion of sentiment classification has appeared in the early 2000s, and has proven to be a challenge that still fuels research. It has practical uses as mentioned before, but the difficulty of the task itself encourages researchers as well.

In this group project, we are interested specifically in the sentiment classification of tweets, both as a whole and as phrases taken out of this corpus.

2 Preprocessing

2.1 Tokenization

Tokenization of tweets is a challenge in itself, as the corpus contains various features not present in texts coming from different sources. They include hashtags (#), tags (@), urls, emoticons, and unconventional abbreviations.

Fortunately, multiple libraries exist to tokenize tweets. The one we have chosen is Natural Language Toolkit [1]. It is capable of much more than just parsing tweets, however we used it only for this purpose.

The output of it is the text split into a list of tokens, which then can be pre-processed.

2.2 Negations

Given a tweet, we have chosen to consider every word that follows a negation to have the opposite sentiment than usual. That is, after a word such as "not", "isn't", "wasn't", etc, and up to and including a token that is followed by punctuation, all tokens have reversed sentiment. More precisely, we construct a regex to capture the negations:

```
regex_negation = "^(not)|[a-zA-Z]+n't$"
```

Each of the words that follow negation then gets prepended with "NOT_" identifier. However, we do not want to prepend the identifier to words such as "but" or to urls, smileys or tags. Those are picked by regex:

```
regex_exception = "^(but|(\W.*)|(http:.*))$"
```

The relation between a word and its negation is such that the weight of the negation is the negative of the weight of the word.

This has to be done obviously before classifying urls, tags, etc., as we replace them with "URLLINK", "USERTAG", or similar identifiers.

2.3 Stemming

To ensure that different form of a word are considered as one, we should to perform stemming on our tokenized tweets. To do this, we attempted to use the Stemming library created by Matt Chaput [2].

It is based on the Porter2 algorithm for stemming English language. The full description of the algorithm can be found here: `http://snowball.tartarus.org/algorithms/english/stemmer.html`

While working on this part, there was a concern that the algorithm was overfitting and truncating too much, and after comparing the scores, there was no significant difference. We decided to leave it out.

2.4 Reading data

Finally we can scan the training files to count how many times each word occurs with each sentiment. The outcome of the preprocessing is a dictionary of words which appear in the corpus, each one associated with an array:

`{'word' : [positive count, neutral count, negative count]}`

3 Classification

The challenge was to find a way of converting the aforementioned dictionary of words and counts for each sentiment, to a dictionary of words and weights.

The aim was to assign a score s to each word, with $s \in [-1, 1]$. From the score we can read off the sentiment as follows:

Given a fixed $\epsilon > 0$, if:

- $s < -\epsilon$ the sentiment is negative
- $-\epsilon < s < \epsilon$ the sentiment is neutral
- $s > \epsilon$ the sentiment is positive

The first approach was a naive mean weight calculated by the formula:

$$\text{weight} = \frac{\text{positive count} - \text{negative count}}{\text{total count}}$$

Where total count means the number of times the word appears in the corpus.

We have also experimented with different weight functions,such as

$$\text{weight} = \frac{(\text{positive count} - \text{negtive count}) \times (\text{total} - \text{neutral count})}{\text{total}^2}$$

However all of the functions we tried gave results very similar (differing by less than 1%) to the results given by the naive formula. Therefore we have chosen to stick to the naive one.

We also had to adjust the interval for neutral sentiment, determined by the value of ϵ.

After training the program on the cleansed training files and running on dev files, we obtained the following results for different values of epsilon:

References

[1] S. Bird et al., *Natural Language Processing with Python*, O'Reilly Media Inc, Sebastopol, CA, 2009

[2] M. Chupat, stemming-1.0, available: `https://pypi.python.org/pypi/stemming/1.0`

ϵ	Training set A	Training set B
0	63.9%	
0.5		

Table 1: Comparison of scores for A and B for a different interval for neutral sentiment