Sentiment Analysis

Arthur Jansen

February 2018

1 Abstract

This report tends to summarize my discoveries of sentiment analysis. I first explain the theory and then present a Python implementation I did to present what can be done with sentiment analysis.

2 Introduction

Sentiment analysis aims to understand the attitude of a person, using artificial intelligence. Even though sentiment analysis can be applied to speeches (to understand if the person is happy, angry, etc.), I focus more on text applications in this report. Natural Language Processing (NLP) is used for this utilization. The most basic task of sentiment analysis on texts is to understand if the overall message of the text is positive or negative. A more complex task is to assess a wider range of emotions.

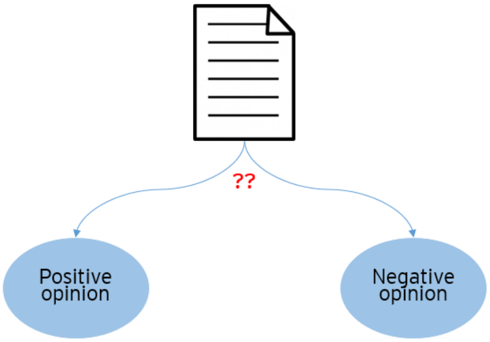

Figure 1: Basic task of sentiment analysis applied on texts is to understand if a text unit is positive or negative.

A common application of sentiment analysis is recommendation systems. For instance, assessing automatically if a review on a e-commerce website is

Rating	Title	Review
3	more like funchuck	Gave this to my dad for a gag gift after direc...
5	Inspiring	I hope a lot of people hear this cd. We need m...
5	The best soundtrack ever to anything.	I'm reading a lot of reviews saying that this ...
4	Chrono Cross OST	The music of Yasunori Misuda is without questi...
5	Too good to be true	Probably the greatest soundtrack in history! U...

Table 1: First five reviews as presented in the dataset They are characterized by their rating, their title and their review in itself.

positive or negative. If we can understand more precisely if a person likes or not something, we can form groups of consumers and better recommend future products.
Another application is to understand the global mindset of a bunch of tweets [1].

3 Dataset

The dataset used in this report contains three millions of Amazon reviews. It was created by. Each review has a score out of five stars. The dataset contains exactly 600,000 occurrences of each score. Each review consists in a title and a text representing the review in itself. Table 1 depicts the first five reviews.

In the following of this section, I will plan to describe the dataset as precisely as possible, especially by trying to characterize the reviews based on their rating. In the next section, I will then use more analytic methods to infer a link between the reviews and their rating.

3.1 Lengths of *title* and *review*

A simple idea to characterize the reviews based on their rating is to calculate the mean length of the title and the text of reviews. It couldn't be too surprising to see longer (or shorter) text reviews associated to a positive or negative rating.

4 Analytic methods

This section is devoted to the introduction of several analytic methods that could be used for sentiment analysis. For each of them, I first explain its concept and then present its results on our dataset.

4.1 N-grams

N-gram is a sequence of n consecutive words from a sentence. An example of bigrams ($n = 2$) is depicted in Figure 2. N-grams are useful for sentiment analysis because the presence of specific groups of words can determine the view of a sentence. For example, more often than not, the bag of words ”$theproductdoesn'twork$” is found in a negative review on Amazon. Therefore, a basic sentiment classifier could be to count the number of n-grams present in positive and negative reviews.

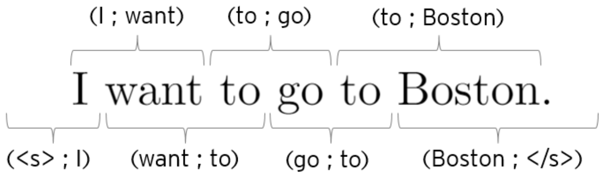

Figure 2: Example of bigrams from a short sentence. Beginning and end of sentence is marked by a special token.

5 Conclusion

Through this report, it has been seen that having two entries for each review (the title and the text of the review) could give more information on the sentiment even though it was sometimes more challenging to find it.

References

[1] Mullen Tony. Introduction to sentiment analysis. University Lecture, 2017.