Advanced Natural Language Processing

Assignment 7

Report by:

Anusha Ramamurthy, Anup Bharadwaj

1. Understanding the problem statement and gathering data: 2 hour

Given the word sentiment analysis, which is probably the most over abused word, our initial step was to understand what sentiment analysis refers to. After reading the papers and more articles on the Internet we felt that Sentiment analysis is way of automation. It can be defined as a programmatic approach to classifying a statement or an article as having a positive or a negative tone. It could also mean identifying the opinion of the author. We would like to think of it as an inference of a reading material, be it a newspaper article or a tweet or a movie review.

With the amount of digital data now available to us it is no longer feasible to employ a hundred people to make inference. So we now have computer programs that are able to identify with some accuracy the sentiment of the material.

2. Our approach was an intuitive one. 4hrs

- Understanding and formulating it took us about 1hr.
- Implementation took us about 2 hrs.
- 1hr to optimize the code to run on a larger dataset.

The best way to understand how the computer should do sentiment analysis lies in the way in which humans have learnt to identify a joke or sarcasm. It lies in the words that are used. "Bad", "Horrible" would easily signify to a child that its a negative tone, the child would then realize that it is being reprimanded.

So all we have to do is build a glossary of word banks? Those that mean to denote positivity and negativity?

This brings us to our python code implementation, this code takes a dataset, identifies all words that have a positive intonation from the "positive"/"negative" label associated with the review. This was our training dataset. Now the algorithm is given a test dataset, which contains no labels, and based on the training set, tries to identify the words that radiate positive or negative context.

We ran this algorithm on movie reviews (small dataset with 1000 reviews) and got an accuracy of about 72%. We used the Naive Bayes Classifier. On a larger dataset (50,000 movie reviews) we had better accuracy of about 83%.

3. Drawbacks of our approach: 1hr

Our classifier learns from training data. Data that was previously labelled by a human to be positive or negative. Hence we would need a different approach for a dataset that has had no human intervention.

Consider the same example of how a child learns to identify sentiment. A bigger part of how the child learns to tell the different, is the variation in the tone of voice a parent uses while reprimanding a child. Similar to an adult identifying sarcasm. When it comes to just two element identification, like "positive" and "negative", we feel that this approach of identifying word similarities and detecting polarity would definitely yeild high accuracy rates. But when we deal with data to detect more complex emotions like sarcasm, I feel our model will fetch lower accuracy results. Why? Well, the whole point of sarcasm lies in the tone of voice used. A human is able to identify a tone of voice even while reading textual data because our mind is able to recreate the tone by looking not only at each individual word but at the construction of the sentence. Our model is blind, in the sense that it can only see unigrams right now. It cannot identify the underlying tone of voice or the structure of the sentense.

4. Additional methods: 2hr

If given 2 weeks for the same, we would like to explore more complex opinion analysis methods like using machine learning models, SVM and Decision Tree Methods. We would also like to understand if bigrams and n-gram methods influence the analysis. Understand how sentence structures influence sentiments. We would also like to explore how constructing sentiment analysis for a different language that has a very different sentence structure compared to English needs a different approach.

Some of the interesting tools we would like to explore are R language and its packages to do sentiment analysis rather than Python Packages. Twitter analysis with the AlchemyApi Twitter analysis tool and its packages.

5. Conclusion: 1hr

To conclude, sentiment analysis seems to be a hot area with diverse applications. We feel that it has the potential to lead to a lot of meaningful social behavioral analysis. Another area that we might consider exploring is "depression emotion

analysis". Analyzing tweets, blogs and forum topics and datasets to identify people who are exhibiting a high level of suicidal tendencies and finding ways to help them through targeted ads to seek counseling. However we are skeptical about the accuracy given the factors we highlighted earlier. Since this is a sensitive issue and might have adverse effects, we should definitely proceed cautiously.