CS 671: Natural Language Processing

Sentiment Analysis in Twitter

Project Proposal

Sakaar Khurana [10627]
Rohit Kumar Jha [11615]

October 21, 2013

1 Introduction

In the past decade, new forms of communication, such as microblogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them. We plan to work on these two tasks and will continue with the one for which we get better results. These two tasks were part of SEMEVAL 2013 challenge. The tasks

- Given a message containing a marked instance of a word or phrase, determine whether that instance is positive, negative or neutral in that context.
- Given a message, classify whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

2 Motivation

Working with these informal text genres presents challenges for natural language processing beyond those typically encountered when working with more traditional text genres, such as newswire data. Tweets and texts are short: a sentence or a headline rather than a document. The language used is very informal, with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations,

such as, RT for "re-tweet" and # hashtags, which are a type of tagging for Twitter messages. How to handle such challenges so as to automatically mine and understand the opinions and sentiments that people are communicating has only very recently been the subject of research (*Jansen et al., 2009*; *Barbosa and Feng, 2010; Bifet and Frank, 2010; Davidov et al., 2010; OâĂŹConnor et al., 2010; Pak and Paroubek, 2010; Tumasjen et al., 2010; Kouloumpis et al., 2011)*.
Another aspect of social media data such as Twitter messages is that it includes rich structured information about the individuals involved in the communication. For example, Twitter maintains information of who follows whom and re-tweets and tags inside of tweets provide discourse information. Modelling such structured information is important because: (i) it can lead to more accurate tools for extracting semantic information, and (ii) because it provides means for empirically studying properties of social interactions (e.g., we can study properties of persuasive language or what properties are associated with influential users).

3 Details

3.1 Approach

We attempt to explore innovative features some of which are listed below and discourse relations to build classifiers which are known to perform well on similar tasks like SVM, MNB(which is known to perform well on text data and is faster than SVMs).

- If a tweet consists of more than one sentences, we give more weightage to sentences coming afterwards.
- We plan to use the hash tags to get idea about the tweets. But the hashtags are like this: # IndiabeatAus # FinallySuccessful and so on. So, we would need to parse these tweets before processing. These hashtags would be structured though not complete sentences. We will keep the weightage of these hastags twice that of the remaining text. It can however be adjusted, depending on results.
- We need to consider abbreviation used commonly. We may even need to prepare a small dictionary for the same manually.
- Incorporate a model to capture the use of idioms, in order to capture sarcasm and other details in the language. One way is to use 4-grams for this as most of the idoms are of length not more than four. But this one requires some more thought and we are still working on it.
- Smileys, again, should have twice the weightage compared to the rest of the text. And the weightage of the smileys increases as we go towards the end of the sentences. (Same idea as in point 1)
- Consider this tweet: "Such a great knock. Team scored this at the loss of just one wicket." Now the problem is that it contains one word "great" and the other "loss",

and so we would get the overall sentiment as neutral. but it is indeed positive. It is important to capture the idea, as to why it is so. The reason is that they say 'loss of "only" one', meaning at a minimal loss. So, if we capture this notion as well, we will get a pretty increase in accuracy. This is something that keeps appearing in texts, inclusing tweets. So, we plan to consider prepositions like "of", "in", "by", etc in vicinity of these sentiment/opinion words. We now plan to study, in detail, the effect of these prepositions.

- We are also planning to incorporate the effect of modifiers like "very", "too", etc

3.2 Data

We currently plan to use these data with labelled tweets as positive, negative or neutral:

- This free data set is for training and testing sentiment analysis algorithms. It consists of 5513 hand-classified tweets. Each tweet was classified with respect to one of four different topics. This has been obtained from the website of Sanders Analytics, a Seattle-based startup focused on data analytics.
  http://www.sananalytics.com/lab/twitter-sentiment/sanders-twitter-0.2.zip

- Sentiment140 Lexicon: The sentiment140 corpus (Go et al., 2009) is a collection of 1.6 million tweets that contain positive and negative emoticons. The tweets are labelled positive or negative according to the emoticon.
  http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

- SEMEVAL 2013 has also provided with around 30000 labelled tweets for the "Contextual Polarity Disambiguation" problem and another 10000 for the "Message Polarity Classification" problem.
  http://www.cs.york.ac.uk/semeval-2013/task2/index.php?id=data

References

[1] S. Mukherjee and P. Bhattacharyya. Sentiment analysis in Twitter with lightweight discourse analysis, December 2012.