CS 144 Ideas Behind the Web
Team: Dai Wei, Doris Xin

Project Plan: Sentiment Analysis on Tweets and Facebook Status Updates and The Effect of Homophily on the Diffusion of Opinions

Introduction

This natural language processing project was largely motivated by the success of the IBM Jeopardy! Challenge as demonstrated through the recently televised Jeopardy! competitions between the Watson system and reigning human champions. Sentiment analysis at both the phrase level and the document level has been the focus of many recent research projects in NLP, and areas of application for such analysis are numerous and varied, ranging from newsgroup flame filtering and informative augmentation of search engine responses to analysis of customer feedback.

In this project, we propose to perform sentiment analysis on the domain of Facebook and Twitter updates because of the challenging nature of this category of texts. As experience tells us, proper grammar and spelling are not expected of texts of this nature, and slang words that have reference to pop culture and whose meanings can vary significantly based on the context are prevalent in messages in this domain. Also, positive comments tend to be straightforward but negative and critical opinions tend to be expressed through sarcastic and sometimes ambiguous language. An additional advantage of studying Facebook and Twitter updates is that after annotating for the positive or negative sentiment associated with each message, we get to explore the effect of the structures of these popular social networks on the diffusion of opinions.

For the data of this project, we would like to construct a network of opinions with each node containing a list of update messages an individual on Facebook or Twitter has posted and the linkage reflecting the friendship relation on Facebook and the following/followed relationship on Twitter. The reason for preserving the network component while fetching the messages mostly lies in the fact that we wish to analyze the diffusion of opinions as affected by the structure of the network, but we also hope to exploit structure in these social networks to help decipher the topic of each message, as locality and homophily has a large bearing on the topics of news to which one is exposed.

As a first step in the analysis, we need to identify the topic of each message. This task is mostly trivial for tweets since we could just parse the hashtags in each message for topic identification. However, Facebook status updates do not provide the luxury of hashtags, and we need to put more thought into the task of topic extraction for these messages. A naive initial approach would be compiling a list of popular topics from news articles and searching for the occurrence of these keywords in each message. As mentioned above, we could also perform topic extraction by looking at the prevalence of certain keywords in connected components in the social networks. We can also make use of existing text parsing systems such as *Lydia* for topic extraction. With minor modifications, we should be able to apply them to our target texts.

We propose to use comments gathered from Yelp! as labeled training data for this project. We're aware of the structural and topical difference between Yelp! comments and Facebook and Twitter updates. However, we believe that Yelp! comments provide a viable source of training data nonetheless for the following reasons:

1. Yelp! comments are a great source of labeled data that requires virtually no manual curation since the rating on each comment serves as an effective indicator of the sentiment expressed in each comment.
2. We expect similarity in the style of language between Yelp! comments and Facebook and Twitter updates due to the large overlap in clientele of these online services.
3. We expect that the meaning of phrase to context relation is roughly preserved across texts generated through online user input and that the Yelp! comments would be able to capture most of the trending phrases. (e.g. "omgomgomg this restaurant was *da bomb*" and "Dude check it out! President Obama's SOTU speech was *da bomb*" both express approval of the subject.)

The feature that distinguishes this project from the status quo is that it attempts to interpret bad English as opposed to dismissing such data as noise. In order to process incorrectly spelled words, slang words, and other computer lingoes, we could try to manually create mappings of letters or words specifically designed for interpreting certain types of nonstandard language(e.g. 3 → e in "th3 cak3 was d3licious"). We could also use these rules to compare the hamming distances between words to map nonstandard spellings onto the correct ones. In addition, we can incorporate an interactive component in which the polarity of nonstandard phrases are manually annotated. People would also be able to group similar phrases together as part of the interactive component (e.g. "cool", "kool", and "kewl" should be considered isomorphic to each other).

To study the effect of network structure on diffusion of opinions, we look at how much correlation there is between clusters seen in the social network and the clusters of positive, negative, and neutral opinions expressed by the individuals in the networks. Because of the fact that we wish to study the distribution of opinions in the network as part of this project, messages containing unpopular topics (sparsely distributed in the network) will be discarded from the data on which we perform sentiment analysis. We perform this analysis on a per topic basis and only focus on popular topics with a significant density in the network. It would also be interesting to see how the topics distribute over the network. We would also like to look at the influence of nodes with high centrality measures on the opinions of of their surrounding community.

Literature Review

There are many established methods for sentiment analysis at the sentence and paragraph level. Mullen and Collier 2004 discussed the application of support vector machines in sentiment analysis with diverse information source. Bang and Lee 2004 applied minimum cuts in graphs to extract the subjective portion of texts they were studying and used machine learning methods to perform sentiment analysis on those snippets of texts only. Wilson et al 2005 discussed categorizing texts into polar and neutral first before determining whether a positive or negative sentiment is expressed through the text. However, Godbole et al 2007 operates on the premise that little neutrality exists in online texts. The followings are a selection of works directly related to the project.

Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens: Automatic Sentiment Analysis in On-line Text. In *Proceedings ELPUB2007 Conference,* 2007.

> This work provides a good survey of various techniques developed in online sentiment analysis. It covers concept of emotion in written text (appraisal theory), various methodologies which can be broadly divided into two groups: (i) symbolic techniques that focuses on the force and direction of individual words (the so-called "bag-of-words" approach), and (ii) machine learning techniques that characterizes vocabularies in context. Based on the survey, Boisy *et al* found that symbolic techniques achieves accuracy lower than 80% and are generally poorer than machine learning methods on movie review sentiment analysis. Among the machine learning methods, they considered three supervised approaches: support vector machine (SVM), naive Bayes multinomial (NBM), and maximum entropy (Maxent). They found that all of them deliver comparable results on various feature extraction (unigrams, bigrams, etc) with high accuracy at 80%~87%.

N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *ICWSM '07*, 2007.

> Godbole *et al*. developed techniques that algorithmically identify large number (hundreds) of adjectives, each with an assigned score of polarity, from around a dozen of seed adjectives. Their methods expand two clusters of adjectives (positive and negative word groups) by recursively querying the synonyms and antonyms from WordNet. Since recursive search quickly connects words from the two clusters, they implemented several precaution measures such as assigning weights which decrease exponentially as the number of hops increases. The confirm that the algorithm-generated adjectives are highly accurate by comparing them to the results of manually picked word lists. It is worth pointing out that this work uses Lydia as the backbone to process large amount of news and blogs.

Alexander Pak, Patrick Paroubek: Twitter as a Corpus for Sentiment Analysis and Opinion

Mining. In *Proceedings of LREC,* 2010.

> Pak *et al* took a naive approach to collect and classify 300000 tweets into three categories: (i) tweets queried with emoticon queries such as “:-)”, “:)”, “=)” indicate happiness and positive emotion (ii) tweets with “:-(”, “:(”, “=(”, “;(” implies dislike or negative opinions, and (iii) tweets posted by newspaper accounts such as “New York Times” are considered objective or neutral. This serves as the training set for naive Bayes multinomial (NBM), which they found to be superior to SVM and CRF (Lafferty et al., 2001) as the classifier to unigrams, binagrams, and trigrams. The result indicates that bigrams provides the best accuracy.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z. and Kellerer, W. (2010) Outtweeting the Twitterers: Predicting Information Cascades in Microblogs, Boston, MA.

> Instead of focusing on the natural language in tweets, Gluba *et al* tracks 15 million URL exchanged among 2.7 million Twitter users. Their data analysis in the cascading of URL uncovers social graphs and other properties on Twitter network. They further formulate a model to predict URL cascading, which accounts for more than half of the URL spread with low false-positive rate.

Timeline

Gather Data for Training and Analysis (1 week)
During the initial phase, we need to collect and curate comments from Yelp! for labeled training data, and updates/tweets from Facebook and Twitter for sentiment and opinion diffusion analysis. Specifically, we would like to download the comments from Yelp! in plain text files that are easily parsable. The updates/tweets from Facebook and Twitter should be downloaded in a format, possibly XML, such that the linkages can be easily interpreted. Features of the messages such as hashtags in tweets should also be preserved by the data format. Parsers for these data files should also be implemented during this stage.

Topic Extraction and Language Processing (3 weeks)
For this phase, we develop techniques for identifying the topic of a message. As mentioned earlier, this is mostly trivial for tweets mainly because of the hashtags. However, this task is nontrivial for Facebook status updates. Yelp! comments also provide a good training set for this task since we can easily query the comments by topic. We can look at employing NLP techniques and searching for specific patterns around topic words, but we can also use the network structures for topic extraction. For instance, a phrase that has a high density in a localized network is likely to be a topic. We can test our techniques on Yelp! comments that were not part of the training set. We’re aware of the structure difference between Facebook status updates and Yelp! comments; therefore, to fully validate the topic extraction techniques,

some manual labeling of Facebook status updates is required. If these techniques fail to extract topics at a high accuracy, we can resort to manually seeding the list of available topics with popular topics and identify the topic of a message by looking for occurrence of these specific topics.

Sentiment Analysis (4 weeks)
This stage is the focus and thus the most time-consuming part of this project. A good deal of the time will probably be spent on studying existing NLP techniques for sentiment analysis by reviewing relevant literature. An appropriate subset of these techniques should be implemented so that their performance can be evaluated on the specific data set for this project. Based on the performance evaluations, we may choose to adhere to the techniques implemented that are proven to be very effective or create new sentiment identification techniques by synthesizing and modifying existing techniques. It is also a good idea to create new techniques and see how they compare with existing ones. At the end of this phase, we should have implemented and evaluated sentiment extraction techniques whose receiver operating characteristics on Facebook and Twitter updates should be comparable with status quo. Manually labeling of Facebook/Twitter updates is once again required in this phase for performance evaluation. If time allows, we will also look at multi-dimensional opinion mining, which is above and beyond the traditional polarity analysis (positive vs. negative comments). In addition to developing and implementing the techniques, we also wish to derive and prove some properties about the techniques and performance guarantees.

Opinion Diffusion Analysis (2 weeks)
We study how network structures play a role in the diffusion of opinions for the last phase of the project. Specifically, we will look at the effect of homophily and network centrality of nodes on the distribution of positive and negative comments on a per topic basis. We rely on the sentiment extraction techniques developed before to create maps showing the distribution of positive and negative comments in the network. We test the hypotheses that homophily in the social networks corresponds to clusters in the opinion distributions and that node with higher centrality measures tend to be surrounded by a higher percentage of nodes sharing the same sentiment as the central node.

End Product

The end product is essentially composed of the deliverables talked about in the Timeline section. Upon successful completion of the project, we should have at least one functional implementation of a sentiment extraction system that performs polarity analysis on Facebook and Twitter updates with relatively high accuracy. In addition, we will have discovered some relationships between the distribution of positive and negative comments and the structure of the social network. As a side product, we will have a well-annotated (through manual curation) database of Facebook and Twitter updates upon successful completion of this project. In the optimal scenario, not only will we have a functional sentiment extraction system, we will also have proven some properties about the techniques used and be able to provide performance bounds on the system through theoretical studies.

Unfortunately, there are many scenarios in which the project could deviate from the planned course. It is possible that we will face obstacles in the very first stage of the project and not be able to acquire a sufficient amount of the desired data for analysis. In the case that Facebook status updates are difficult to obtain, we will probably have to drop the Facebook portion of the project and work primarily with tweets, which are readily available through exisitng sentiment analysis projects on tweets. It is also conceivable that processing slangs and nonstandard English is a greater challenge than we expect and that topic extraction becomes a difficult task. As mentioned in the Timeline, if this were the case, we have a simplified topic extraction system in place that will probably guarantee decent performance so that we can move onto the next stage. Since manual curation of the data is required for this project, the quantity and quality of the useble data may be limited by the manpower available for curation. An insufficient amount of data may lead to large underestimation of the performance of the techniques. Without a working sentiment extraction system, we will not be able to study diffusion of opinions in the two social networks used in this project.

References

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (ACL '04).

N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In ICWSM ’07, 2007.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (HLT '05).

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP*.

Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens: Automatic Sentiment Analysis in On-line Text. In *Proceedings ELPUB2007 Conference,* 2007.

Alexander Pak, Patrick Paroubek: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC,* 2010.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z. and Kellerer, W. Outtweeting the Twitterers: Predicting Information Cascades in Microblogs, Boston, MA., 2010.

Naaman, M., Boase, J. and Lai, C. Is it really about me?: Message content in social awareness streams, *CSCW 10*, 189-192, 2010.