Peer-to-Peer Insult Detection in Online Forums

Submitted By:
Priya Goyal (10535)
Gaganpreet Singh Kalra (10258)
IIT Kanpur

Advisor:
Prof. Amitabha Mukerjee
IIT Kanpur

Problem and Motivation

Now days, most people are connected across the world through online blogs, news groups and social network sites to express their opinions, share and receive knowledge. This communication involves people from many different cultures, communities and different parts of world. However, sometimes people may use language that is considered inappropriate by others and may hurt others feelings. Besides, it may lead to frustration among the users searching for particular information on some specific site because some people take it as fun to use personal attacking and insulting messages. While the *Terms of Service* for sites e.g. Facebook, Twitter etc. prohibit posting content that is unlawful, abusive and harassing, users' posts are only partially filtered for some particular collection of offensive words. Also, while some others like newsgroups, YouTube etc. provide users with the facility to mark a comment as insulting/ inappropriate, they are prone to collusion and are highly misused. There is no existing classifier that identifies insult speech directed towards a participant of the conversation. We aim at building such a classifier in this project.

Related Work

Various attempts have been made to classify the insulting comments in on-line forums. The work by Ellen Spertus [1] used static dictionary approach and defined some patterns based on socio-linguistic observation to build a feature vector for training. This work suffers from high false positive rates and low coverage. Another work by Altaf Mahmud et. al [2] tries to differentiate between insult and factive statements by parsing the sentences and using the semantic rules. This works doesn't distinguish between the insults directed to non-participant and participant of conversation. Also the rules and seed words approaches are rigid and lack generality because of flexibility of conversation. The work by Razavi et. al [3] makes use of Insulting and Abusing language dictionary on top of bag-of words features in their proposed three-level classification machine learning model. Another recent effort by Carolyn P. Rose et. al [4] in classifying offensive tweets builds topical features and lexicon features and uses various machine learning approach.

However, besides the limited approach of using seed words and pattern matching, these works also do not distinguish between the insult directed towards the people participating in blog/forum conversation and non-participants such as celebrities, public figures etc. Comments which contain profanity or racial slurs may not necessarily be insulting to other person. We aim at building an efficient classifier that detects peer-to-peer insults and we adopt the machine learning approach entirely.

Dataset

The train data typically consists of labels (insult/ non-insult), timestamp of comment and the comment. The size of train data is around 7000 comments collected from kaggle.com site. The dataset is nearly balanced with 3000 insulting comments. The test data consists of 2500 comments from the forum.

Approach

We use supervised learning methods for this one class classification problem. We first process the data by removing html tags, "\\n", \\xc2" etc. and by changing the most common used words like "u" to "you", "da" to "the" etc. other patterns that were found in the dataset. Then we normalize our dataset along with

the badwords file available online [5] as people circumvent the dictionary that may otherwise put bias on results. We then do stemming of words like “retarded” to “retard”, “embarrassing” to “embarrass” etc. to avoid dealing with large data. For feature extraction, we plan to use tf-idf scores. Based on our observation of dataset, we find a lot of insulting comments involved phrases like “you are an XXX” where XXX is derogatory word. We plan to use this second-person rule for our training. We also plan to use the n-grams model and machine learning techniques like logistic regression, SVM for model building. Then the test data is tested by combining the classifiers built.

REFERENCES

1. Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence, pages 1058–1065.
2. Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan, 2008. Detecting flames and insults in text. In Proceedings of the Sixth International Conference on Natural Language Processing.
3. Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In Proceedings of the 23rd Canadian Conference on Artificial Intelligence, pages 16–27.
4. Xiang, G., Hong, J., & Rosé, C. P. (2012). Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus, Proceedings of The 21st ACM Conference on Information and Knowledge Management, Sheraton, Maui Hawaii, Oct. 29- Nov. 2, 2012.
5. For badwords file:
   http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/