1. Introduction

Opinion mining and sentiment analysis are sub-disciplines in the area of natural language processing, which search into analyzing and extracting people’s opinion, sentiment and attitude from written text.

Most research work has been focusing on extracting and analyzing a set of pieces from the text, these include:

* Identifying the subjectivity of a sentence
* Identifying the polarity of a subjective sentence either positive or negative
* Identifying the opinion holder
* Identifying the aspect of the opinion

Advanced classification looks beyond the two classes of positive and negative, it will classify the opinion into emotion classes such as happy and sad.

2. Sentiment Analysis in Twitter

Twitter is an online social networking site that provides a micro blogging service. Users can post and read text updates that are limited to 140 characters called tweets. An average of 58 million tweets per day is posted[[1]](#footnote-2); due to the huge number of users and user generated content twitter it is considered one of the leading social networks.

The users can tweet about anything anytime of the day, opinions can be expressed and exchanged freely, and hence it is a rich source of opinionated text. Sentiment analysis for entities in twitter became the interest of many researchers.

Twitter is different sentiment analysis source in a way that tweets contain a lot of noisy data like emoticons and URLs, and users usually use abbreviated expressions in order to maintain space within 140 character limit.

The aim of this paper is to investigate and summarize the techniques used so far by researchers to extract sentiment from tweets and the methods used to exploit noisy data in tweets to the advantage of research. The rest of this article is arranged as follow: section 2 review the available corpora and lexicons for twitter and the techniques used to collect the data, section 3 summarize the most used techniques for sentiment analysis in tweets

3. Twitter corpora and lexicons

Available access to a corpus of annotated tweets is a necessity for the purpose of training and learning. Collecting and labeling such data is a laborious work, Here is a description of the automatic techniques proposed by researches to collect a big number of tweets with annotation without the need for a manual annotation of tweets.

The main idea of automatic annotation of tweets is to make use of some elements in the tweet which could be considered as noisy labels, such as emoticons and hash tags.

3.1 Emoticons

One approach is to treat emoticons in tweets as sentiment indicator [Go et al., 2009; Bifet,2010; pak et al.,2010; Davidov et al., 2010; Speriosu et al., 2011]. Emoticons is a representation of facial expressions frequently used in social networking websites, each emoticon show a certain state of emotion and one emotion can be expressed using different kinds of emoticons. In the work of [Read, 2005] emoticons proved the be a good method to reduce the dependency of time and domain in machine learning t techniques. In twitter data collection there are basically two types of emoticons used:

- Happy emoticons to express positive sentiment:  
 :) , =) , :D

- Sad emoticons to express negative sentiment:  
:( , =( , :'(

The collection of tweets is done by the help of Twitter API, [Go et al., 2009] query twitter API using a list of topics to get the tweets and from the query results then they annotated based on the emoticons that the tweet contains. [Pak et al., 2010] followed the same approach for collecting data, In addition they collected a set of tweets that represent objective posts by querying tweets from popular newspapers such as New York times.

3.2 Hash tags

In twitter, hash tags is used as a method to categorize tweets or specify the topic of the tweet, and it is created by proceeding the topic word with a hash '#UTM'. Hash tags can be created and utilized by any registered user, adding a hash tag to the tweet increase the tweet visibility.

[Wang et al., 2011] Categorized hash tags into three types:

- Topic hash tags that serves as user annotated coarse topics like '#Obama'

- Sentiment hash tags that serves as a sentiment label for the tweet like '#love or #hate'

- Sentiment-Topic hash tags where the topical word and the sentiment regarding this topic appear together in the hash tag without separating space, example of this kind of hash tag: '#iloveobama'

In regard to the hash tag part of the tweet Some hash tags constitute an essential part of the tweet, such that if it is deleted the tweet will lost the meaning, as an example:[ #ImNiceUntil you eat my food without my permission!]. Other has tags only label the tweet such as [i love my chat room in Facebook #windows]

In [Davidov at la, 2010] they categorized the hash tags into five categories: Strong sentiment (#sucks) - most likely sentiment (#notcute) - context dependent sentiment (#shoutsout) - focused sentiment (#tmobilesucks) - no sentiment (#obama). Their dataset was formalized from tweets that lays within the first or second category.

Collecting the dataset based on hash tags is the same as with emoticons, twitter API is used to query the hash tag.

3.3 Lexical seed and rule based

Lexicons are used in a number of techniques of sentiment analysis. a sentiment lexicon contains the sentiment words and the label of each word either positive or negative

3.4 Preprocessing

Automatic collection of dataset can result in many unwanted or repeated tweets, plus one tweet can contain abbreviated text or URLs that may or may not help in the learning process. So, preprocessing the data before learning is an important step for better results. The most used preprocessing techniques are as follow:

remove non English tweets (Davidov et al., 2010)

delete retweets and repeated tweets (Zhang et la., 2011)

Remove URLs, Usernames and Special twitter characters such as 'RT' which is the abbreviation of retweet (Pak et al., 2010; Zhang et la., 2011)

replace emoticons with its sentiment indication by searching through an emoticon dictionary. Such as replacing the emoticon ':)' with the 'positive' token (Agarwal et al., 2011;

replace hash tags with a token (Kouloumpis et al., 2011; Davidov et al., 2010

replace URls with a special tag (Go et al., 2009; Kouloumpis et al., 2011; Agarwal et al., 2011; Davidov et al., 2010

Replace mentions '@' or username with a special tag (Go et al., 2009; Kouloumpis et al., 2011; Agarwal et al., 2011; Davidov et al., 2010

remove repeated letters replace the occurrence of more than two letters in one word by two or one occurrence, example: the word (huuuuuungry) will be (Huungry) (Go et al., 2009; Kouloumpis et al., 2011; Agarwal et al., 2011;

extend abbreviated expressions to the actual meaning such as: OMG -> oh my GOD (Kouloumpis et al., 2011; Agarwal et al., 2011; Zhang et la., 2011)

tokenization separating the words based on space and punctuation marks (Kouloumpis et al., 2011; Pak et al., 2010)

removing stopwords (Pak et al., 2010

part-of-speech tagging (Kouloumpis et al., 2011; Zhang et la., 2011)

Replace negation words with a special tag such as 'NOT' (Agarwal et al., 2011)

4. Models of analysis

Tweet level – Sentence Level

Aspect Level

Hash tag Level

[Barbosa, 2010] Used other available twitter sentiment tools to collect the training data.

The problems specific to twitter and what kind of information to extract – subjectivity – polarity or aspect

Although twitter limit the number of written characters in one tweets to 140 Char, We cannot take for a fact that one

tweet will express either positive or negative sentiment. One tweet can contain the both positive and negative opinion such as {TODO}

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Paper  Process | Go et al., 2009 | Davidov et al., 2010 | Pak et al., 2010 | Zhang et la., 2011 | Agarwal et al., 2011 | Kouloumpis et al., 2011 |
| remove non English tweets |  | X |  |  |  |  |
| POS tagging |  |  |  | X |  |  |
| Remove URLs, Usernames and RT |  |  | X | X |  |  |
| Replace URL | X | X |  |  | X | X |
| Replace Emoticons |  |  |  |  | X |  |
| Replace Hashtags |  | X |  |  |  | X |
| Replace Username | X | X |  |  | X | X |
| Remove repeated letters | X |  |  |  | X | X |
| extend abbreviations |  |  |  | X | X | X |
| tokenization |  |  | X |  |  | X |
| removing stopwords |  |  | X |  |  |  |
| Replace negation words |  |  |  |  | X | X |

Table 1: Preprocessing Techniques

Preprocessing

Datasets

Preprocessing to achieve higher accuracy with the learning process, delete noisy data that can lead to false classification.

Techniques

Machine learning based

Features divided as mentioned in abbasie

The last section of machine learning mention the techniques used to handle negation, conditional sentences or comparative sentences.

When mentioning part of speech tagger shows that some papers came to the conclusion that POS does not have that much contribution to the accuracy of the final result or classification. Then mention the paper that proposes a special POS tagger for twitter where it introduces new parts that are specific to twitter.

Lexicon based

At the end of the lexicon based papers mention hp paper that combine the two previous approaches machine learning and lexical based.

Graph based

Ontology based

Available corpus and lexicons

Models of analysis:

Tweet level – Sentence Level

Aspect Level

Hash tag Level

One section can talk about corpus collection and the methods used in it

Noisy data and preprocessing and how some papers use these noisy data as a label for the tweet

Manual labeling

Emoticons as labels

Hashtags as labels

Lexical seed and rule based

The last section can talk about **evaluation** and the usual way to evaluate and what other ways like evaluating unbalanced data

Sarcasm in twitter

Try to connect between the techniques already used and the unsolved problems

Twitter features:

Some features in twitter is considered as noisy data that need to be pre-processed before use

Some can be exploited in the process of collecting the training set such as hash tags and emoticons

Others can be used as features for the classifier to learn from

References

|  |
| --- |
| Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 36-44). Association for Computational Linguistics. |
| Bifet, A., & Frank, E. (2010, January). Sentiment knowledge discovery in twitter streaming data. In *Discovery Science* (pp. 1-15). Springer Berlin Heidelberg. |
| Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012* (pp. 508-524). Springer Berlin Heidelberg. |
| Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ... & Smith, N. A. (2010). *Part-of-speech tagging for twitter: Annotation, features, and experiments*. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE. |
| Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30-38). Association for Computational Linguistics. |
| Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1-12. |
| Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC* (Vol. 2010). |
| Read, J. (2005, June). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop* (pp. 43-48). Association for Computational Linguistics. |
| Saif, H., He, Y., & Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. In *The 2nd Workshop on Making Sense of Microposts*. |
| Kouloumpis, E., Wilson, T., & Moore, J. (2011, May). Twitter sentiment analysis: The good the bad and the omg. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 538-541). |
| Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011, July). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP* (pp. 53-63). Association for Computational Linguistics. |
| Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 241-249). Association for Computational Linguistics. |

1. http://www.statisticbrain.com/twitter-statistics/ [↑](#footnote-ref-2)