International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 05, May-2015

Extended Comprehensive Sentiment Analysis for Informal Opinion Text

Mansi Kataria
ME Student, Department of Computer Engineering,
GEC, Sector-28,
Gandhinagar, India.

S. M. Shah
Associate Professor, Department of Computer Engineering,
GEC, Sector-28,
Gandhinagar, India.

***Abstract* - There has been a remarkable increase in use of E-commerce websites, the World Wide Web can now be seen as a repository of reviews and opinions from users spread across various websites and networks. Comprehensive feedback about the product can help the company know what the users liked and disliked about their product. People tend to adapt modern writing styles like misspelled words, abbreviations, concatenated words and emoticons, considering these can increase the accuracy of sentiment analysis. Moreover people tend to use different words for a particular feature of the product. Thus, identifying frequent nouns and noun phrases automatically will help classify more number of reviews and also help in identifying any new feature of the product that is being talked about.**

Keywords - Sentiment analysis, Informal text, Noun phrases, Product reviews.

I. INTRODUCTION

Today an enormous volume of data is being generated and making any sense out of that data is a tedious task. Lots of research is done recently to automate this task of sentiment analysis of the data. Sentiment analyses refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. By combining more than one sentiment analysis techniques we can increase the accuracy of sentiment analysis. To provide a comprehensive feedback, it requires detailed sentiment analysis of the data under consideration. There are three levels of sentiment analysis: Document level, Sentence level and Aspect level. Out of these the most detailed sentiment analysis is the Aspect level sentiment analysis, i.e. for every feature of the product in case of product reviews.

Now a days people are also rapidly developing modern writing habits that include misspelled words, abbreviations, emoticons, etc. informal texts. If these are taken into consideration, it can increase the accuracy of sentiment analysis. This type of informal text can be seen especially where there is a constraint on the number of words that can be entered, for e.g. SMS, Tweets. Due to lack of space people tend to use abbreviations and slang.

Comprehensive sentiment analysis means that it includes.everything i.e. Detailed sentiment analysis for all text types (formal and informal), and also displaying the results such that it will answer all questions that a user or production company might have. Data visualisation APIs are used for displaying such results.

For Aspect-level sentiment analysis, the system should be able to identify the features about the product on its own. Automating the feature extraction process helps identifying a new feature and we do not need to manually add it to the list of features whenever a new feature comes

II. RELATED WORK

Summarizing the results of aspect level sentiment analysis using data visualisation API is necessary in order to give the user a detailed result for analysing the product. The work of Kherwa Pooja et al. [1] presents such an approach. Aspects of the product are identified manually and SentiWordNet [12] is used for sentiment analysis. Then the summary of result is displayed using Google-o-meter. I.Hemalatha et al. proposed a method for sentiment analysis of product reviews from twitter in [2]. Their proposed method includes preprocessing of informal text by removing URLs, removing questions, filtering words with repeated letters, removing special characters and re-tweets. This leads to efficient sentiment analysis. Vibha Soni et al. proposed a method of sentiment analysis in [3] where features of the product are identified and SentiWordNet is used for sentiment classification. The summarized result is displayed using a chart. Jmal et al. proposed a system in [4] where features of the product are identified automatically by the system and emoticons are considered to identify the polarity of the review statement about a feature. Aditya Joshi et al. [5] developed a system called C-Feel-It for sentiment analysis in microblogs. Here also preprocessing of informal text is done by handling extensions in words and normalization of chat words in the microblog. Words used in chat/Internet language that are common in tweets are not present in the lexical resources [5]. Subhabrata Mukherjee et al. presented TwiSent [6] that again does sentiment analysis for general tweets (not product reviews). TwiSent, inspired from C-Feel-IT [5], is a Twitter based sentiment analysis system. However, TwiSent is an enhanced version of their rule based system with specialized modules to tackle Twitter spam, text normalization and entity specific sentiment analysis [6]. David Garcia et al. in [19] uses SentiStrength [10] as a lexical resource. The lexicon used is an extension to SentiStrength, with the addition of

emotional terms from ANEW [9] dataset. Thus, informal text can also be dealt with by combining two or more lexical resources in this way.

Hanen Ameur et al. in [7] presents a method for constructing dynamic dictionary (lexicon) specific to the domain under consideration so as to create a lexicon covering several sentiment words. Aminu Muhammad et al. in [8] uses a combination of two knowledge sources Generic Lexicon and Domain specific lexicon. Also separate scores are maintained for acronyms such as lol, ugh, etc. This way there is an improvement in accuracy of sentiment analysis and informal text is also taken care of. Similarly, existing lexical resources can be modified as in [9].

Feature extraction is done by finding out noun phrases in [4]. Whereas in [15], whenever a new product is searched, its features are extracted and stored in json database. Reference [16] gives a review about feature extraction techniques proposed and used in various research papers.

III. PROPOSED WORK

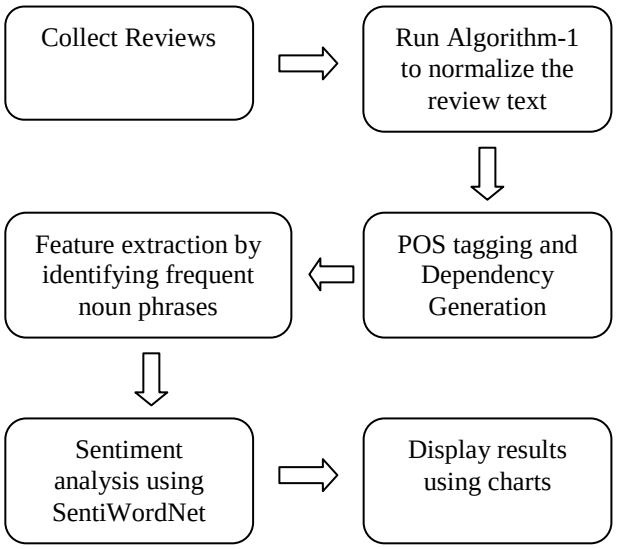

Figure 1:Different steps of our method of sentiment classification

Each of the above steps are described below:

Step-1: Product Reviews are collected from various E-commerce sites including twitter because there is a greater possibility of finding chat language/internet slang language on twitter because of writing space constraint.

Step-2: Algorithm-1 is proposed to convert the informal text to correct English text. It is based on two well-known spelling correction algorithms namely Peter Norvig's Spell checking algorithm [18] and Soundex algorithm [19]. The Peter Norvig's algorithm is slightly modified and is combined with Soundex algorithm. Peter Norvig's algorithm is a simpler and smaller algorithm made similar to Google's spelling correction algorithm. Whenever we type something in Google, it immediately says *Did you mean spelling?* Peter Norvig developed, in less than a page of code, a toy spelling corrector that achieves 80 or 90% accuracy at a processing speed of at least 10 words per second [18]. This algorithm gives the correct word for any input word based on Edit Distance and probability of the word in training dataset. Edit distance means the number of insertions, deletions and transformations required to convert one word to another. Whereas soundex algorithm is a phonetic algorithm and encode the homophones to the same representation so that they can be matched despite spelling mistakes. The idea to combine soundex algorithm with Peter Norvig's algorithm came from the fact that reviewers while using chat language in their reviews tend to play around with vowels. Vowels are dropped, interchanged, etc. either intentionally or by mistake. Soundex algorithm drops the vowels and encode the remaining letters, so when a word is played around with vowels it will still match the correct word. The algorithm is follows:

```
Consider S is a set of words from our language model starting with same letter as s.
/*remove extra characters from the word*/
/*let the extra character be c*/
if(s contains more than two c)
        replace the sequence of c by 'cc'
end if
/*get W' subset of S based on Levenshtein distance*/
for all s' in S
        if(Levenshtein_distance(s,s')<=3)
                Copy word s' from S to W'
        End if
End for
/*get W'' subset of W' based on soundex code*/
for all w' in W'
        calculate soundex core of w'
end for
From list of words W'' select the word considering three factors in decreasing order of priority:
    ⇒ Matching soundex code
    ⇒ High probability
    ⇒ Low Levenstein_distance
This is our correct word for s.
```

Algorithm 1

Step-3: Stanford NLP parser can be used to tag the parts of speech in the sentence and generate dependencies. Dependencies in a sentence shows the relations between words in a sentence. After this, only the relevant set of dependencies are kept and others are ignored as in [1].

Step-4: As the sentences are already POS tagged in step-3, in this step we can find out the nouns and noun phrases. Then calculate the frequency of each of these, set a threshold value and keep only the nouns and noun phrases with frequency above the threshold value. There after a similarity measure is made to find which group of nouns are used to talk about a single feature of the product. This method is quite similar to the one used in [4].

Step-5:After identifying the features, we need to do sentiment classification using SentiWordNet. The method used for sentiment classification is similar to the one used in [1].

Step-6: In order to give the customer a better idea of the result of our analysis, data visualization tools are used to display the sentiment for each feature of the product along with the overall sentiment. This type of representation can be done using Google chats API, JAVA Swing and AWT classes, etc. This step helps in answering all the questions that a person may have regarding the product in a graphical way.

V. Conclusion

As the word 'Comprehensive' means everything about something, in our research we have tried to include everything from collection of reviews to display of results using visualization techniques in sentiment analysis of product reviews. Improvements can be made to the proposed algorithm for spelling correction so that other types of errors in spelling can be dealt with. This will help in more accurate classification of the reviews. Also a Spam Filter technique can be included so that review spam are identified. This will help create a complete comprehensive sentiment analysis system. Our research helped us take a step towards comprehensive sentiment analysis.

References

(1) Kherwa, Pooja, Arjit Sachdeva, Dhruv Mahajan, Nishtha Pande, and Prashast Kumar Singh. "An approach towards comprehensive sentimental data analysis and opinion mining". In Advance Computing Conference (IACC), 2014 IEEE International, pp. 606-612. IEEE, 2014.
(2) Hemalatha, I., GP Saradhi Varma, and A. Govardhan, "Preprocessing the Informal Text for efficient Sentiment Analysis". International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 1, Issue 2, July – August 2012.(pp. 58-61).
(3) Soni, Vibha, and Meenakshi R. Patel, "Unsupervised Opinion Mining From Text Reviews Using SentiWordNet". International Journal of Computer Trends and Technology (IJCTT) – volume 11 number 5 – May 2014. (pp. 234-238).
(4) Jmal, Jihene, and Rim Faiz, "Customer review summarization approach using Twitter and SentiWordNet". In Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, p. 33. ACM, 2013.
(5) Joshi, Aditya, A. R. Balamurali, Pushpak Bhattacharyya, and Rajat Mohanty, "C-Feel-It: a sentiment analyzer for micro-blogs". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations, pp. 127-132. Association for Computational Linguistics, 2011.
(6) Mukherjee, Subhabrata, Akshat Malu, Balamurali AR, and Pushpak Bhattacharyya. "TwiSent: a multistage system for analyzing sentiment in twitter". In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 2531-2534. ACM, 2012.
(7) Ameur, Hanen, and Salma Jamoussi, "Dynamic construction of dictionaries for sentiment classification". In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on, pp. 896-903. IEEE, 2013.
(8) Muhammad, Aminu, Nirmalie Wiratunga, Robert Lothian, and Richard Glassey "Domain-Based Lexicon Enhancement for Sentiment Analysis". In SMA@ BCS-SGAI, pp. 7-18. 2013.
(9) Arup Nielsen, Finn. "A new anew: Evaluation of a word list for sentiment analysis in microblogs". arXiv preprint arXiv:1103.2903, 2011.
(10) Thelwall Mike. "Heart and soul: Sentiment strength detection in the social web with sentistrength". Cyberemotions (2013): pp.(1-14).
(11) Araújo, Matheus, Pollyanna Gonçalves, Meeyoung Cha, and Fabrício Benevenuto. "ifeel: A system that compares and combines sentiment analysis methods." In Proceedings of the companion publication of the 23rd international conference on World wide web companion, pp. 75-78. International World Wide Web Conferences Steering Committee, 2014.
(12) Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In LREC, vol. 10, pp. 2200-2204. 2010.
(13) Chalothorn, Tawunrat, and Jeremy Ellman, "Sentiment Analysis: State of the Art", Proc. of the Intl. Conf. on Advances in Computer and Electronics Technology -- ACET 2013 pp.(21-25).
(14) Chalothorn, Tawunrat, and Jeremy Ellman. "Sentiment Analysis Of Web Forums: Comparison Between SentiWordNet And SentiStrength". The 4th International Conference on Computer Technology and Development (ICCTD 2012). 24-25 November 2012, 2012.
(15) Singh, Prashast Kumar, Arjit Sachdeva, Dhruv Mahajan, Nishtha Pande, and Amit Sharma. "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites". In Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-, pp. 329-335. IEEE, 2014.
(16) Asghar, Muhammad Zubair, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi. "A Review of Feature Extraction in Sentiment Analysis". Journal of Basic and Applied Scientific Research 4, no. 3 (2014): 181-186.
(17) Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis". Foundations and trends in information retrieval 2, no. 1-2 (2008): 1-135.
(18) Norvig, Peter. "How to write a spelling corrector." URL: http://norvig.com/spell-correct. html (2007).
(19) Mitton Roger. "Spellchecking by computer." Journal of the Simplified Spelling Society 20, no. 1 (1996): 4-11.