International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

Procedure of Opinion Mining and Sentiment Analysis: A Study

Rushabh Shah[Á*] and Bhoomit Patel[Á]

[Á]Information Technology,Dwarkadas J.Sanghvi College of Engineering,Vile Parle(W), Mumbai-400056, India

Accepted 02 Dec 2014, Available online 10 Dec 2014, **Vol.4, No.6 (Dec 2014)**

Abstract

This paper covers all the essential details one must know about sentiment mining. It provides information on recent trends, applications of sentiment mining, different fields where it is used and also lot of useful information on the current research work being carried out in this area of data mining. Also, the basic workflow of the sentiment analysis process has been explained extraordinarily. Further, this paper also exemplifies the challenges and the future research being planned in the field of sentiment and opinion mining.

Keywords: *Data Mining; Data Pre-processing; Analysis; Opinion; Tweets.*

1. Introduction

As we know internet has provided us with variety of means to flourish all groups of industries. All the social medias like twitter, MySpace, Linkedin, Facebook, YouTube and many others have gained so much reputation that they cannot be ignored. Internet offers effective means to communicate and share one's opinion. It simply allows people to frame links with their colleagues, friends and families. It allows people to share all kinds of information and use different types of services available such as sharing blogs, reviews, etc.

Opinions regarding almost all the global entities are available on the internet. We find many blogs dedicated to particular topics like sports, news, marketing, finance, education, history, science and many more. Opinions are expressed by the people in the form of natural language.

Social networking sites can easily provide one with all the information required to take a particular decision, for example, buying any item A from shopping sites. Social media is an excellent channel to put forward one's opinion in front of the world. The data available in order to mine the opinion from it is magnificent. There have been many research projects based on the analysis of sentiments expressed on social media. Sentiment analysis poses newer and various challenges to gather information from the text in natural language. The region of Sentiment analysis aims to understand all the opinions expressed in natural language and categorize them. Sentiment analysis is carried out on review sites and social medias like twitter where tweets gives us more accurate and varied opinions of the people from all over the world which can be about latest cellular phone like Iphone6. The reviews regarding a product would definitely affect the buyer's decision.

The key challenge faced by the researchers in analyzing these reviews from the internet is that they are in the form of natural language. Processing of natural language is implicitly problematic , gathering information from the unstructured reviews is even more problematic. In this paper we have discussed most of the key challenges faced in opinion mining. We have included the flow of sentiment analysis for understanding of the concept. The paper includes ongoing research and future scope of opining In this paper we have considered twitter as social media.

2. Recent Trends

Sentiment analysis is not a novel research theme. The use of automation in sentiment analysis has increased significantly in the past couple of years. Natural Language Processing, Machine Learning and Opinion mining are few streams of computer science on which the research theme is dependent. Nowadays, lots and lots of data is available and can be adopted for use. For making any important decisions, devising business strategies it is a necessity to analyse enormous amount of data available from various sources. Social media is one of the upcoming sources providing indefinite data which is used for sentiment analysis. But the data obtained from social media is disorganized. Obsolete content analysis methods focused on predicting topics. Since past few years opinions, emotions and sentiments are some qualities which are reflected by the content from the social media. The volume of data obtained from social media is gigantic making it more complex to analyze. As a result, decrease in interest of semantic-based application and inclination towards statistics and visualization is observed.

3. Applications

Sentiment mining covers a vast range of applications in several fields. These applications assist in making sense of

*Corresponding author: **Rushabh Shah**

hundreds of applications. Sentiment mining works diversely from the traditional survey methods and depends on listening in spite of asking which depicts more accurate reality. The main domain of applications implemented and managed by sentiment mining are as follows:

- Applications to review related websites

  It has the same capabilities as a review-related search engine and acts as an alternative to sites such as Opinions that collects information in the form of reviews and feedback. With such applications it becomes possible to summarize user reviews, fix blunders in user ratings and provide evidence that proves that user ratings was biased or those ratings need some correction.

- Applications as sub-component technology

  Sentiment analysis is having a potential role in supporting technologies used for other systems. Detecting 'flames' in email or other means of communication, augmenting to recommendation systems which avoids recommending items getting lot of negative response, detecting web pages that includes sensitive information and averts displaying ads on those pages are some of the examples of application being executed in this field.

- Applications in business and government intelligence

  Sentiment mining is extremely important when business intelligence is the factor one is focusing on. In business and government intelligence, sentiment mining mainly handles reputation management, public relations and monitoring sources responsible for increment in negative or hostile communications. Extracting information helps organizations to develop better business strategies, find answers to their decline and failure, review their products based on people's comments or tweets which would all help any organization walk on the path of success.

Other applications include areas like politics, question answering, summarizing important points, improving extraction by discarding petty information, citation analysis, strong holding human-computer interaction etc.

4. Workflow

4.1. Extraction

In the extraction phase of Sentiment mining, social media acts as a source of data. In order to explain this process easily further details are in resemblance with twitter. In twitter, number of users gives their reviews by posting messages which are called as tweets. These tweets depict the sentiments of the users. The process of sentiment mining is basically analyzing this data and converting it into knowledge. Following are few fundamental characteristics observed in the data while performing extraction:

- The length of the twitter message is limited to 140 characters.
- Moreover, we observe the presence of spelling errors and informal or cyber slang in these messages.
- The amount of data available is copious and as most of the twitter messages are available in public domain it can be used for the purpose of sentiment mining.

Data extracted from the social media like twitter is updated very frequently. Therefore, it helps to give the feeling of real time representation of the sentiments. In order to obtain the data on run-time an internet bot can be used known as web crawler. A web crawler browses through the World Wide Web in organized manner to index the web pages. It is one of the many fundamental components which constitute web search engines. The indexing of the web pages grants the user to issue queries and get the required pages as per the query issued.

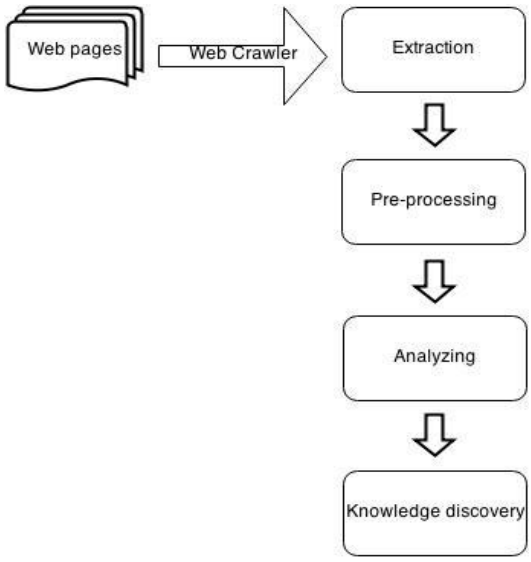

Fig.1 Working of sentiment analysis

4.2. Pre-Processing

As the name suggests, in this phase of sentiment analysis processing of the data is carried out. In the pre-processing stage, the extracted data is cleaned as it contains large amount of noise before sending the text for analyzing. The extracted text contains lot of grammatical errors as the text is of limited length. Pre-processing of the data is necessary and it is a crucial part as one needs to make sure that the unnecessary part of the text is removed and the relevant part of the text which stores the sentiment of the user is not removed. Following are few techniques explained in brief, keeping in mind twitter as the social media used, which are generally used in order to draw information for sentiment mining.

- Supplanting emoticon

  In twitter as one is restricted to post his or her views by using only 140 characters, emotions is proved to be an easy way to depict one's sentiments. Moreover,

Emotions assist in deciding the polarity of the text. For example, a SMILE keyword can be used to supersede few similar emotions such as :D , :] , =) , etc.

- Uppercase and Lowercase Identification

  It is often observed that in order to express strong sentiments (like anger e.g. GET OUT) one uses all the characters in uppercase. It is generally used to show the intensity of the emotions. This is considered as an indicator to decide the polarity of the text and is known as e-shouting. We observe inconsistent casing (e.g. TwITteR) in texts on social media. It is necessary to make sure that there is consistency in casing of the texts.

- Extraction of URL

  In order to share extra content due to the limitation of the tweet, many tweets consist of URL which is an extension to the posts. The information obtained from the URL basically supports the sentiments which are expressed in tweets. But the cost of crawling is very expensive. Therefore a compact equivalent class <URL> can be used consisting of all URLs.

- Pointer Detection: We have observed that many people uses '@' ahead of the users name in order to point to other users and '#' is used by users to tag twitter posts related to some category. Similar to URL, <USER> and <HASHTAG> can be used to supplant this portion of the tweets.

- Punctuations: In order to avoid the usage of grammar rules and to give stress on the emotions expressed in the tweets excessive usage of the punctuation is done by the users. (E.g. Hurrah!- to show excitement). In this technique of pre-processing stage, removal of punctuation marks which are not relevant is done.

- Compression of words: Informal language is generally used by the users of twitter and many users stretch the words in order to highlight stress on those words which clearly describes emotions. For example, 'Air is soooooooo necessary' simply depicts the idea of necessity of that thing. The term 'soooooooo' simply bears the idea of importance of the 'Air'. Therefore, we can reduce it to shorter sequence like 'soo'. It is not stored as usual term 'so' so that it can be used to differentiate.

4.3. Analysis

Sentiment analysis is carried out on the data which is obtained after the pre-processing stage. From the sentiments contained in the data, the number of repetitions observed in the tweets and the location of the tweets is also analyzed. It is a vital stage in sentiment mining. The tweets are mined for different keywords describing emotions and are rated in this stage. If during the tenure of analysis a existing keyword is seen in a sentence, it is scored and is added to the score of the entire text to decide the polarity of the text.

During the stage of sentiment analysis it is better analyze on the basis of a scale and not merely on binary judgment because co-relation compared the predicted value to the target value which makes it better than precision. But most of the algorithms for sentiment analysis use basic terms to review and express opinions about a service or product. So in this stage, proliferating contexts, cultural factors, linguistic differences are taken into consideration to get a precise result and the complexity of deciding whether a sting of text is a pro or a con sentiment gets simpler and efficient.

4.3. Knowledge Discovery

To find the opinion of the people with respect to any particular occurrence, it is essential to store the data which is related to the event. Once the polarity of the sentiments is known it can be used to generate statistical graphs and charts. The knowledge gathered from these electronic texts from the web when shown in graphs would aid the individuals in making decision as it would show the polarity of the sentiments of the individuals and to what extent it can be followed by referring to the graphs.

After all these stages are completed, the process of sentiment mining is successfully executed.

5. Current Research

Nowadays, profiles like data mining, data analysis, business analysis have become really prominent. For any organization to stay alive in corporate world, it has become necessary to make judicious decisions based on a tremendous amount of data from ample of sources. But the quantity and quality of data available from web sources such as blogs, social media and discussion forums are copious with sentiments, opinions, therefore, current research is directed towards the area of opinion mining.

It is essential to compose a system which can identify the sentiment or opinion from the text available and easily classify the sentiments or opinion. In order to identify sentiments, current research is hunting down methods for reduction of human efforts to perform operations while analyzing the content. Classification of sentiments with the help of sentiments which is already known by using the pre-existing dictionary of words obtained from the electronic text from the web. Visual mapping of bipolar or conflicting opinions is also something n which the current research in concentrating.

Also, the focus is on ameliorating the precision of algorithms for opinion detection, identifying the policies for analyzing opinionated materials and finalizing the experts for sentiment analysis who are highly experienced and capable of working smoothly with large amounts of data.

But, the main subject on whom the current research is working is improving the precision of algorithms for opinion detection.

6. Key Challenges

Opinion mining has become really important as nowadays for devising even smaller strategies, corpuses of data has to be analyzed. So solutions for opinion and sentiment mining are evolving at a great pace and reducing the burden of human shoulders. There are many techniques available for sentiment analysis. But still it's very difficult to say which technique works best because each technique has its own issues and certain challenges.

Mainly there are two techniques used for opinion mining.

1) Lexicon based and 2) Learning based

Lexicon based techniques involves high precision but on the other hand gives low recall. Also, another issue in this technique is that lexicons aren't available in all the languages. Learning based techniques makes use of labeled examples to classify text. But it requires learning training as well as training dataset which becomes as issue too. Another technique-syntactic technique does yield god results but the n it isn't language independent. Few other challenges that erupt during the process of classify text in opinion mining have dependency on factors like:

- Value for n-when n-gram framework is used, choosing a higher value of n will degrade the performance.
- Occurrence of word-the number of occurrence of a word should be at the most 2-3 times for sentiment analysis.
- Features to be used-Deciding what features must be used is also a challenge as the list of features returned by tokenization contains few irrelevant features. Hence it becomes necessary to select relevant features which will determine the precision of a classifier.
- People are habitual to using slang and casual language on social websites which it makes really difficult to predict people's opinion. So to eradicate such issues methods must be developed and existing ones should be modified to adapt to the kind of language used on social websites.

Now the term "spam messages" or "fake reviews" have become really common and used on a large scale on social websites. This creates a hurdle in the process of sentiment mining. So one of the biggest challenges is to identify such spam messages and fake reviews which can mainly be done through the comparison of qualitative with summary reviews. Also, identifying duplicates, detecting outliers and knowing the reputation of the reviewer is to be kept in mind while implementing opinion mining. There are limitations in collaborative filtering which is responsible for identifying most famous concepts and suggest some out of the box thinking. Another challenge is the risk of filter bubble, where combination of automated content analysis with behavioral analysis proves to be very effective but eventually deviates selection of useful opinions making user unaware of content that is different from what he expects in some manner. Integrating opinion with implicit data and behaviour to validate data and provide analysis beyond the opinion expressed is another common challenge. Asymmetry in available opinion mining software and the need for continuous improvement in the usability and user-friendliness of opinion mining software and other tools is another key challenge in the field of sentiment and opinion mining. Skewness in the dataset that is responsible for impacting recall is another challenge in sentiment mining.

Conclusion

The explosion of usage of social networking sites in order to give one's reviews has helped a lot to produce and use varied technologies to mine people's opinions and sentiments. As it involves natural language processing, Sentiment mining arises as a challenging field with many hindrances. It has varied diversity of applications that could prove to be advantageous to many fields such as marketing, business analytics, knowledge bases and so on. To understand texts as human is the key challenge of this field when it comes to machine's ability. It is very important to extract knowledge from the opinions which are expressed on the social networking sites for many companies and institutions, whether it is in terms of public mood, or feedback of particular product. In this paper we have covered all the challenges faced in the current research work. We have analysed the flow of the process of sentiment analysis along with detailed techniques and explanation of the stages in the process.

Many other challenges are there like sarcasm detection for which natural language processing can be used. Similarly there are many other research work in progress and soon will be unveiled in coming years to overcome the current problems and enable smooth functioning eliminating the challenges.

Future Research

Currently there are many significant matters which seek attention. After all these requirements are accomplished, it will become possible to divert attention on the flaws, shortcomings in the current methods, algorithms and devising solutions and strategies to reduce them and make the process of sentiment analysis simpler and more efficient. The future research work can be generalized into either short term or long term research.

Short-term

- Bipolar evaluation of opinions
- Visual representation
- Multilingual collection of writings for reference
- Real-time and dynamic sentiment mining
- Opinion mining across different platforms
- Audiovisual opinion mining
- Algorithms for learning machines
- Algorithms for recommending opinion and comment

Long-term

- Non-bipolar evaluation of opinions
- Detecting irony automatically

- Autonomous machine learning and artificial intelligence
- Developing usable tools for citizens to let them carry out opinion mining

To overcome the problem of skewness in the dataset, techniques such as undersampling and oversampling are being developed. AdaptiveBoost is being used in conjunction with classifiers to eradicate petty entries in the training set which would apparently help in improving recall rates. Classification of international words and foreign expressions is another area where the future research would focus on.

Acknowledgement

We would like to thank Prof. Arjun Jaiswal for giving us an opportunity to work and provide us a helping hand. We would also like to thank our honorable principal Dr. Hari Vasudevan of D. J. Sanghvi College of Engineering and Dr. Abhijit Joshi, Head of Department of Information Technology, for giving us the facilities and providing us with a propitious environment for working in college. We would also like to thank S.V.K.M for encouraging us in such co-curricular activities.

References

Akshi Kumar and Teeja Mary Sebastian (July 2012) Sentiment Analysis on Twitter. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3.

Bo Pang and Lillian Lee (2008) Opinion mining and sentiment analysis. Foundations of Information Retrieval, Vol 2, Nos. 12, 1-135.

David Osimo and Francesco Mureddu (2012) Research Challenge on Opinion Mining and sentiment analysis. http://www.w3.org/2012/06/pmod/opinionmining.pdf),

Mark Kantrowitz (2000) Method and apparatus for analyzing affect and emotion in text. U.S. Patent 6622140, 2003. Patent filed in November

Balakrishnan Gokulakrishnan , Pavalanathan Priyanthan , Thiruchittampalam Ragavan ,Nadarajah Prasath, AShehan Perera (2012) Opinion Mining and Sentiment Analysis on a Twitter Data Stream. The International Conference on Advances in ICT for Emerging Regions - ICTer, 182-188

Emma Haddia, Xiaohui Liua, Yong Shib (2013)The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science 17, 26 – 32.

Walaa Medhat , Ahmed Hassan , Hoda Korashy (2014) Sentiment analysis algorithms and applications:A survey. Ain Shams Engineering Journal.

Mita K. Dalal andMukesh A. Zaveri (2013) Semisupervised Learning Based Opinion Summarization and Classification for Online Product Reviews. Hindawi Publishing Corporation, Applied Computational Intelligence and Soft Computing, Article ID 910706.

Lalita Sharma, Shweta Shukla. Classification of Web Blog Mining for Movie Review. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958.

Meishan Hu, Aixin Sun and Ee-Peng Lim (2007) Comments-oriented blog summarization by sentence-extraction. In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management(CIKM), 901-904,. ISBN 78-1-59593-803-9, Post paper.

Charlotta Engstrom (2004) Topic dependence in sentiment classification. Master's thesis, University of Cambridge.

Soo-Min Kim and Eduard Hovy (2004) Determining the sentiment of opinions. In Proceedings of the International Conference on Computational Linguistics (COLING).