

An Study and Analysis of Opinion Mining and Sentiment analysis

Akshata Shirish Nayak, Roshan Fernandes

Abstract— The outbursts of social media has created millions of opportunities for people or web users to state or express their opinions, but it is a challenging task when it comes to bringing out the meaning or some sense from the opinions. Opinion mining is a natural language processing for finding or knowing the viewpoint of the people or user about a particular topic under discussion or products. Opinion mining involves forming of a system to extract and analyze the view point or stand of different users about the product made on social media and review sites like blogs, comments, tweets or reviews. Opinion mining can be employed in many ways depending on domain or user interests in different fields for different purpose, due to which analyzing the opinions has become very essential.

Index Terms—Opinion mining, Natural language processing, Knowledge extraction.

I. INTRODUCTION

The cheap access of Internet to millions of people and the technology centric generations has contributed to generation of terabyte of data which are primarily in the form of unstructured text. Most of the time the actual reviews are so long that it is impossible to read even a few of them and deduce concrete knowledge from it. In these cases, customers may naturally gravitate to reading a few reviews in order to form a decision regarding the product and he/she only gets a biased view of the product. Similarly, the manufacturer also wants to feel the pulse of the people who are using their product in order to develop marketing plans and product placement in the highly competitive market. The approach could be to use intelligent techniques to collate the information extracted from various data sources into a semantically related structure and provide with a visualization tool which can help the users at multiple levels of complexity.

Today, great amount of information are available on web. As part of the effort to better organize this information for users, the problem of automatic text categorization is been addressed. However, recent the rapid growth in on-line discussion groups and review sites where a major characteristic of the posted comments are the sentiments, or overall viewpoint or their stand towards the subject matter for ex: Is a movie review positive or negative is extracted.

Opinion mining also known as Sentiment analysis refers to a kind of natural Language processing, analyzes the documents so as to identify and extract subjective data. In

general, opinion mining aims to determine the stand or viewpoint of a reviewer in regard to a topic or the overall polarity the reviews or comments.

II. LITERATURE REVIEW

With the rapid proliferation of blogs, wikis, and social networking sites, people are expressing their thoughts more freely, publicly, and frequently than ever before. Users opine on medical topics to mobile companies their stocks and from different product merits demerits to the movie reviews. The ability to leverage the vast amount of user-generated content on these and similar sites for commercial and social benefit is supported via the development of an emerging technology called opinion mining.[1]

A growing number of recent studies have focused on the economic values of reviews, finding the relationship between the sales and performance of products and their reviews [2], [3], [4], [5]. Since what the general public thinks of a product can no doubt influence how well it sells, understanding the opinions and sentiments expressed in the relevant reviews is of high importance, because collectively these reviews reflect the wisdom of crowds (what the general public think) and can be a very good indicator of the products future sales performance.

In this paper [6] authors proposed a technique based on association rule mining to extract product features. Their main idea is that people often use the same words when they comment on the same product features. Depending on if the frequent item sets of nouns in reviews are found then they are likely to be product features and the infrequent are less likely to be product features. But, in fact, the infrequent item may also be features, which may offer more information. Their work only find the features that many people focus on, which not what we want to do. The infrequent features also are very import for people to make choice. The method proposed in this paper uses opinion words to extract the corresponding features. The relationship between opinions and features is used to remove the noise to improve the precision.

In this paper [7] authors proposed a novel algorithm called Double Propagation. It is a state-of-the-art unsupervised technique for solving the problem. Their primary idea is that opinion words are usually associated with features in some ways. Thus, opinion words can be recognized by identified features, and features can be identified by known opinion words. So the extracted opinion words and features are utilized to identify new opinion words and new features, which are used again to extract more opinion words and

features. This propagation or bootstrapping process ends when no more opinion words or features can be found. The biggest advantage of the method is that it requires no additional resources except an initial seed opinion lexicon, which is readily available. It mainly extracts noun features, and works well for medium-size corpora. But for large corpora, this method can introduce a great deal of noise (low precision), and for small corpora, it can miss important features.

III. OPINION MINING APPROACHES

Study the evolution of sentiment analysis research can be done by the analytical tokens, or building blocks and the implicit information associated with those tokens. We can group the existing approaches into four main types they are keyword spotting approach, lexical affinity method, statistical methods, and concept-based techniques.[8].

A. Keyword spotting

This approach makes use of unambiguous affect words such as happy, sad, afraid, and bored to classify the text into different categories. "Today was a great day" as being affectively positive, it is likely to assign the same classification taking an example of sentence like "Today wasn't a great day at all." Keyword spotting also relies on the presence of actual affect of words and not for a sentence.

Sometimes, a sentence conveys through their meaning rather than affect adjective words. For ex the following sentence "My husband just wants to take custody of my children away from me" states strong emotions, but uses no affect keywords, and therefore is ineffective. Lexical affinity is slightly more sophisticated than keyword spotting.

B. Lexical affinity

This approach not just detects obvious affect words, but also assigns arbitrary words a probable "affinity" to particular emotions. For example, lexical affinity might assign the word "accident" a 75-percent probability of indicating a negative effect, as in "car accident" or "hurt by accident." This approach usually trains probability from linguistic corpora. Although it often outperforms pure keyword spotting, there are two main problems with this approach. First, negated sentences (I avoided an accident) and sentences with other meanings (I met my friend by accident) trick lexical affinity, because they operate solely on the word level. Second, lexical affinity is often biased towards a category, directed by the linguistic corpora's source. Due to which it becomes difficult to build and reuse the model for different systems.

C. Statistical methods

This approach is popular for affect text classification. Researchers use statistical methods on projects, by using a machine-learning algorithm for a large training corpus of affectively annotated texts, the system will learn combination of things as in keyword spotting and lexical affinity. Generally, statistical methods are semantically weak, which means that individually with the exception of obvious affect keywords a statistical model's other lexical or co occurrence

elements have little predictive value. Due to which statistical text classifiers give accurate results only when they are given sufficiently large amount of data as input. So, these methods will affectively classify the reviews or opinion text on large for ex page or paragraph level, they don't fare well when it comes to smaller texts such as clauses or sentences.

D. Concept-based approaches

To accomplish semantic text analysis methods such as semantic networks and web ontology are used. This helps the approach to identify the affective and conceptual data related to natural language opinions. In this approach instead of blindly using the matches of keywords and counts of word co-occurrence by using the large semantic knowledge dictionaries, and rather depend on the actual meaning associated with natural language contexts. Concept-based approaches can identify minute expressed sentiments when compared to purely syntactical techniques. Concept-based approaches can analyze multi-word expressions that don't explicitly convey emotion. The concept-based approach relies heavily on the depth and breadth of the knowledge bases it uses. Without a comprehensive resource that encompasses human knowledge, an opinion mining system will have difficulty grasping the semantics of natural language text.

IV. DIFFERENT LEVELS OF OPINION MINING

A. Document level Opinion Mining

The basic data unit is a one document of opinionated text [9] where in document level classification deals with classifying to a single or one review for that entire document. One of the situation where document level opinion mining doesn't give good results is in forums blogs etc where there is a comparison made between two different topics in such a case document level mining can classify to wrong single review. Therefore classification depending on subjective and objective becomes very important.

B. Sentence level Opinion Mining

In sentence level Opinion Mining, each sentence is parsed and the polarity is been derived. The same methods of document level classification can be applied to the sentence level classification as well but subjective and objective sentences [10] must be identified. The objective sentences contain the facts of the product or topics such as price specifications and so on where as in subjective sentences it will have user opinions which can determine the sentiments. Then the subjective sentences are classified to positive or negative classes.

C. Phrase level Opinion Mining

The phrase level sentiment classification is a more sophisticated approach to opinion mining. The phrases are scanned for the opinion words and then the phrase level opinion mining is applied to them. This method also does not work well in all the cases where the negation words and sentences are far apart it can identify only when negation words occur locally [11].

V. OPINION MINING APPLICATIONS

Now that highly accurate opinion information is becoming available commercially, the demand for such services by corporations, governments, nonprofits, and individuals is increasing dramatically.

A. Corporations

Opinion mining methods allow such companies to quickly conduct analyses of online, user-generated comments to determine how at least a segment of the population views a given brand or product. Pharmaceutical companies have a well-grounded interest in tracking what people are saying about the drugs they produce.

Are patients having adverse side effects from the drug?. Manufacturers need to understand who likes and dislikes their products and what features are viewed positively or negatively.

B. Government

In addition to knowing what their citizens are thinking, governments want to understand the impact of their actions. Political parties need public opinion information to determine how their candidates are faring.

C. Individuals

Individuals can also benefit from information companies and insurance companies could easily allow users to search for doctors in their geographic vicinity not only by specialist, but also by scores quantifying the opinions expressed about them. The same is true for lawyers, plumbers, car mechanics, and other service providers.

VI. OPEN SOURCE TOOLS

A variety of open-source text-analytics tools used for natural-language processing such as information extraction and classification can also be applied for opinion mining. Tools are listed below:

A. *Ling Pipe* - It is an open source tool used for linguistic processing of data with respect to extraction, grouping and classification, etc. It is the most widely used open source NLP tool-kit. This tool is considered for its speed, stability, and scalability.[12]

B. *OpenNLP* - Performs the most common and necessary NLP tasks, such as chunking and co-reference resolution, named entity extraction, POS tagging.[13]

C. *Stanford Part-of-Speech (POS) Tagger and parser* - Is used for parsing the sentences and part of speech tagging .[14]

D. *NLTK* - The natural language tool-kit is a tool for parsing, clustering and researching classification.[15]

E. *Opinion Finder* Opinion Finder aims in finding and to mark various aspects of subjectivity in subjective sentences.[16]

VII. CONCLUSION

The Web has changed from read-only to read-write. This evolution created enthusiastic users interacting and sharing through social networks, online communities, blogs, wikis, and other collaborative media. Collective knowledge has spread throughout the Web, particularly in areas related to everyday life, such as commerce, tourism, education, and health. Despite significant progress, however, opinion mining and sentiment analysis are still finding their own voice as new interdisciplinary fields.

ACKNOWLEDGMENT

Our heartfelt gratitude to the Department of Computer Science and Engineering, NMAM Institute of Technology, Nitte for providing us the laboratory facility to conduct this experiment. We also thank for the support of teaching and non teaching staff of this Institution.

REFERENCES

- [1] V.S. Subrahmanian. **Mining Online Opinions** ,Computer,IEEE ,Volume:42 , Issue: 7,July 2009.
- [2] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. **The Predictive Power of Online Chatter** Proc. 11th ACM SIGKDD Intl Conf. Knowledge Discovery in Data Mining (KDD), pp. 78-87, 2005.
- [3] A. Ghose and P.G. Ipeirotis. **Designing Novel Review Ranking Systems:Predicting the Usefulness and Impact of Reviews**, Proc. Ninth Intl Conf. Electronic Commerce (ICEC), pp. 303-310, 2007.
- [4] Y. Liu, X. Huang, A. An, and X. Yu. **ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs**, Proc. 30th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 607- 614, 2007.
- [5] Y. Liu, X. Yu, X. Huang, and A. An. **Blog Data Mining: The Predictive Power of Sentiments**, Data Mining for Business Applications, pp. 183-195, Springer, 2009.
- [6] Minqing Hu and Bing Liu . **Mining Opinion Features in Customer Reviews**, AAAI'04 Proceedings of the 19th national conference on Artificial intelligence, 2004.
- [7] Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. **Expanding domain sentiment lexicon through double propagation**. in Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-2009).
- [8] Erik Cambria, Bjorn Schuller, Yunqing Xia. **New Avenues in Opinion Mining and Sentiment Analysis** ,Intelligent Systems, IEEE,Volume:28 ,Issue:2March-April 2013
- [9] Raisa Varghese, Jayasree. **A Survey on Sentiment Analysis and Opinion Mining**, International Journal of Research in Engineering and Technology (IJRET), Vol 2 Issue 11 Nov 2013.
- [10] Nidhi Mishra et al. **Classification of Opinion Mining Techniques**, **International Journal of Computer Applications**, Vol 56, No 13,Oct 2012Pg No 1-6.
- [11] G.Angulakshmi, Dr.R.ManickaChezian. **An Analysis on Opinion Mining: Techniques and Tools** , International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014
- [12]<http://alias-i.com/lingpipe/>
- [13]<http://opennlp.apache.org/>
- [14]<http://nlp.stanford.edu/software/tagger.shtml>
- [15] <http://www.nltk.org/>
- [16] <http://code.google.com/p/opinionfinder/>

Akshata Shirish Nayak
PG Scholar
Computer Science and Engineering
NMAM Institute of Technology
Nitte, India

Roshan Fernandes
Associate Professor
Computer Science and Engineering
NMAM Institute of Technology
Nitte, India