A Brief Survey Paper on Sentiment Analysis

Archana Kalia
Lecturer, Information Technology Department,
VPM Polytechnic, Thane.

Abstract:-Sentiment Analysis (SA) is a current topic of research in data mining field. SA is the process of computing of opinions, sentiments and subjectivity of text. This survey paper gives a comprehensive overview of this field. Out of many, three topics of SA applications are investigated and presented briefly in this survey. This survey presents a clear idea of SA techniques and with brief details. The main contributions of this paper include the sophisticated categorizations of the methods considered and the illustration of the recent trend of research in the sentiment analysis and its contribution in related areas.

Keywords:- Sentiment analysis, Opinion mining, Document level SA, Sentence level SA, Aspect level SA.

I. INTRODUCTION

Sentiment Analysis (SA) or Opinion Mining (OM) is the arithmetic, algebraic and algorithmic study of people’s opinions, attitudes and emotions toward an entity. The entity can be represented as individuals, events or topics. SA and OM express a mutual meaning. However, some researchers stated that OM and SA have slightly different convictions. Opinion Mining is the process of extraction and analysis of people’s opinion about an entity while Sentiment Analysis identifies the sentiment expressed in a text or document and then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then categorize their polarity. There are three main classification levels in SA: Document-level, Sentence-level, and Aspect level SA.

Document-level SA analyze a text document and declare it as positive or negative opinion or sentiment. It considers the whole document as a basic information unit or entity.

Sentence-level SA tries to analyze sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions.

Aspect-level SA analyze the above sentiment with respect to the specific aspects of entities. It performs fine-grained analysis and directly looks at the opinion. The goal of this level of analysis is to discover sentiments on aspects of items.

II. OPINION MINING OF MOVIE REVIEWS: AN DOCUMENT LEVEL APPROACH

In this paper an Opinion Mining System is considered. It is termed as “Document based Sentiment Orientation System” which is based on unsupervised approach that determines the sentiment orientation of documents. Sentiment orientation determines the polarity of documents. It classifies the documents as positive and negative. This approach helps the users in decision making by providing the summary of total number of positive and negative documents.

The above approach determines the opinion words from the documents and classifies the corresponding polarity of the documents. The unsupervised dictionary-based technique is used in this system. WordNet is used as a dictionary to determine the opinion words and their synonyms and antonyms.

There are various websites available on the web which contains movie reviews. These movie reviews are collected from different different websites which contain the user and critic reviews.
All the collected reviews are applied to the proposed system which classifies the reviews as positive, negative and neutral. Final results are presented in graphical charts.

III. SENTIMENT ANALYSIS ON TWITTER DATA USING SUPPORT VECTOR MACHINE: A SENTENCE LEVEL APPROACH

Sentence level approach uses different machine learning classifier Naïve Bayes and Support vector machine. The feature is used is unigram and bigrams. The process starts by getting the tweets from twitter, then passes by each tweet and labels it as positive, or negative. After that the features in each tweet will be extracted and represented in a feature vector. Then, these feature vectors will be used in the training phase of the classifier.

Before determining the polarity of the collected reviews, preprocessing of the collected reviews are necessary to get the cleaned reviews. Pre-processed reviews are applied as input.

Unlike the binary classification problem, it was suggested that opinion subjectivity and expresser credibility should also be taken into consideration. A framework was proposed that provides a compact numeric summarization of opinions on micro-blogs platforms.

The above process identified and extracted the topics mentioned in the opinions associated with the queries of users, and then classified the opinions using SVM.

It was found out that the consideration of user credibility and opinion subjectivity is essential for aggregating micro-blog opinions.

It was proved that the above mechanism can effectively discover market intelligence (MI) for supporting decision-makers by establishing a monitoring system to track external opinions on different aspects of a business in real-time.

IV. SENTIMENT ANALYSIS ON TWEETS ABOUT DIABETES: AN ASPECT-LEVEL APPROACH

In recent years, some methods of sentiment analysis have been developed for the health domain. An aspect-level sentiment analysis method based on ontologies in the diabetes domain was proposed. The sentiment of the aspects was calculated by considering the words around the aspect which are obtained through N-gram methods (N-gram after, N-gram before, and N-gram around).

To evaluate the effectiveness of our method, a corpus from Twitter was obtained, which has been manually labeled at aspect level as positive, negative, or neutral. The experimental results show that the best result was obtained through the N-gram around method with a precision of 81.93%, a recall of 81.13%, and an -measure of 81.24%.

The proposed sentiment classification approach is divided into three main components:

- Preprocessing module,
- Semantic annotation module, and
- Sentiment classification. Figure below shows the architecture of the system
A. Preprocessing Module:

This module consists in the preprocessing of the corpus to clean and correct the text.

a) It involves five processes. The first process, called normalization, consists of three main tasks.
   - The special characters that do not provide important information were removed.
   - Correction of spelling errors
   - Replace the abbreviations and shorthand notations by their expansions which are not identified by the Hunspell dictionary.

b) The second step, known as tokenization, consists in dividing text into a sequence of tokens, which roughly correspond to “words.”

c) The third step involves assembling the tokenized text into sentences.

d) The fourth step consists in processing a sequence of words and assigning a lexical category to each word.

e) The fifth step refers to the process of mapping words to their base form.

- **Semantic Annotation Module:** This module involves the detection of aspects by means of the semantic annotation technique. The semantic annotation is carried out through of a natural language processing (NLP) tool, namely, Stanford NLP, and in accordance with domain ontology.

- **Sentiment Classification:** This module calculates the polarity of each aspect found on the SentiWordNet (SWN) lexicon.

In this segment, the sentiments of the aspects of each tweet are obtained. This process has been carried out using the “N-gram after,” “N-gram before,” and “N-gram around” methods. N-gram technique collects the number of words which are near of the aspect which are then considered for the sentiment analysis.

In the current work the N-gram of high information gain feature extracted from the tweets is combined with the sentiment lexicon to train the classifier and evaluate the performance of predicted result.

V. COMPARATIVE STUDY

After surveying the above three types of reviews, a comparative study can be stated based on different levels of sentiment analysis.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Document level</th>
<th>Sentence level</th>
<th>Aspect level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>To determine if the document as a whole has a negative or positive sentiment</td>
<td>To determine if the document as a whole opinion has a negative or positive sentiment</td>
<td>It is the most fine-grained analysis. This analysis finds a target for each opinion, instead of focusing on language units, like sentences, documents or paragraph.</td>
</tr>
<tr>
<td>Data set</td>
<td>Movie Reviews</td>
<td>Twitter data</td>
<td>Tweets about health issues like diabetes</td>
</tr>
<tr>
<td>Technique used</td>
<td>Unsupervised learning dictionary-based technique such as WordNet.</td>
<td>Supervised learning such as Support vector method using height accuracy level.</td>
<td>Lesion based approach using N-gram method.</td>
</tr>
<tr>
<td>Accuracy level</td>
<td>Accuracy level is low.</td>
<td>Accuracy level is more than that of document level.</td>
<td>Highest accuracy level.</td>
</tr>
</tbody>
</table>

Result Based Summary:

- With the help of above technique, the polarity of the document is determined.
- Polarity is determined on the basis of majority of opinion words, if the number of positive words is more, then the polarity of document is positive otherwise the polarity is negative and if the number of positive and negative words is equal then the document shows the neutral polarity.
- SVM has defined input and output format. Input is a vector space and output is 0 or positive/negative. Text document in original form are not suitable for learning. They are transformed into format which matches input of machine learning algorithms input. For this preprocessing on text documents is carried out. Then we carry out transformation.
- Each word will correspond to one dimension and identical words to same dimension. SVM has been proved one of the powerful learning algorithms for text categorization.
- N-gram parser has been used to read a source text file, divides it into content to sentences on newline, and reads the number of grams in the combination with the result of the compression unit. In N-gram parser, we use five kinds of gram to these for unigram, bigram, trigram, four grams, and five grams. The result obtained at compression unit will be used by N-gram parser to read text further.
VI. PROCESS OF SUMMARIZATION

Apart from analyzing and extracting opinion information from individual documents, the process of summarization involves aggregating and representing sentiment information drawn from an individual document or from a collection of documents.

The process of summarization is of different following types:

A. Single-Document Sentiment Summaries:
When user might desire an at-a-glance presentation of the main points made in a single review, it is considered as single-document sentiment summaries.

B. Multi Document Sentiment Summaries:
The automatic determination of marketsentiment or the majority “leaning” of an entire body of investors, from the individual remarks of those investors is a type of multi-document opinion-oriented summarization.

C. Generic Summarization:
Generic summarization assumes that the audience that reads the summary is a general one. It determines the appropriateness of including a phrase or a sentence into the summary only based on the information contained in the input documents.

D. Focused Summarization
While focused summarization, targets at generating summaries for specific information of interest, especially for the information requested by users.

Specifically, a query-focused summarization system usually takes a question asked by a user, and 10 then generates a summary with respect to the query, ignoring all other content from the original document(s).

E. Extractive Summarization:
The most prominent multi-document summarization approaches have been extractive summarization methods, where sentences from the original documents are selected for inclusion in the final summary. Extractive methods have been popular mainly because they are relatively simple to construct, since the problem can be converted to a sentence selection task and the output summary does not suffer from ungrammaticality.

F. Abstractive Summarization
When people write summaries, they tend to abstract the content and seldom use entire sentences taken closely from the original documents. If human summaries are compared with the input documents, we can observe several operations on how humans use and modify the input content: sentence compression, information fusion, paraphrasing, and generation.

Therefore, summarization research has moved towards the area of abstractive summarization. Abstract-based methods are often designed to approximate how human construct summaries.

VII. BROADER EFFECTS OF SENTIMENT ANALYSIS

Sentiment-Analysis technologies have some of the larger effects that hamper the existence of opinion-oriented information-access.

A. Privacy:
It should be a matter of concern that applications that gather data about people’s preferences can trigger concerns about privacy violations.

B. Manipulation:
Since sentiment-analysis technologies allow users to consult many people who are unknown to them, hence that it is difficult for users to evaluate the trustworthiness of those people they are consulting. Thus, sentiment analysis technologies might probably make it easier for users to be mis-led by destructive entities, a problem that originator of such systems might wish to prevent.

C. Economic Impact:
Many customers who are swept by online reviews say that these reviews significantly influence their purchasing decisions.

D. Interactions with Word of Mouth (WOM):
Some studies point out that the number of reviews, positive or negative, may simply reflect “Word of mouth”, so that in some cases, what is really the underlying correlative (if any) of economic impact is not the amount of positive feedback but merely the amount of feedback in total.

VIII. CONCLUSION

The process of sentiment analysis involves various text analytics technique and accepted opinion mining process to determine and find out subjective information from given entity.

Thus sentiment analysis process determine how a certain person or group reacts to a subject matter they are being referred to. They react because they are either interested or involved. And, these reactions get stored in their social media accounts which make social media as one of the leading platforms in the internet where anyone can basically do opinion mining.
Maximum businesses are mostly benefitted from sentiment analysis now-a-days. Few individuals refer to it as social site to analysis it since it also typically analyzes the ongoing activities which are ongoing in these major social networking sites. The paper shows that businesses can solely count positive and negative reviews of their brands. Thus it also helps them to measure their overall performance, especially on their online presence.

On the other hand, certain people can also get a lot from opinion mining. They are making a class or identification for themselves or just trying to know anything that regards to them. Actors, celebrities, famous writers and all other popular individuals can definitely benefit from the idea of sentiment analysis technique. They can simply learn how to inspire the common public, how people show their reaction (negatively and positively) to any recent move they make and which of it stimulate people's attitude towards them.

REFERENCES

[1]. Sentiment Classification with Convolutional Neural Networks: an Experimental Study on a Large-scale Chinese Conversation Corpus By Lei Zhang Email: lei.zhang@xiaoi.com, Chengcai Chen Email: arlene.chen@xiaoi.com
[3]. Twitter Sentiment Analysis with Deep Convolutional Neural Networks Aliaksei Severyn, Google Inc. aseveryn@gmail.com
[4]. Alessandro Moschitti, Qatar Computing Research Institute, moschitti@qf.org.qa
[5]. Survey on sentiment analysis of movie reviews Neha Nehra M.E (CSE), L.J Institute of Engineering and Technology, Gujarat Technological University, Gujarat, India
[6]. Applications of Deep Learning to Sentiment Analysis of Movie Reviews,
[7]. Houshmand Shirani-Mehr, Department of Management Science & Engineering Stanford University hshirani@stanford.edu
[8]. A Literature Review on Opinion Mining and Sentiment Analysis Md. Daiyan1, Dr. S. K. Tiwari2, Manish Kumar3, M. Aftab Alam4
[9]. 1,3Research Scholar, Institute of Computer Science & IT, Magadh University, Bodh Gaya, Bihar 2 Associate Professor, Department Of Mathematics, Magadh University, Bodh Gaya, Bihar 4Research Scholar, Department of Computer Science & IT, Senghania University, Pacheri Bari, Rajasthan.
[10]. Convolutional Neural Networks for Sentence Classification Yoon Kim New York University yhk255@nyu.edu
[11]. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts C’icero Nogueira dos Santos, Brazilian Research Lab IBM Research cicerons@br.ibm.com, Ma’ira Gatti, Brazilian Research Lab IBM Research mairacg@br.ibm.com