An Overview of Sentiment Analysis

Abhijit Bhosale#1, Abhishek Kulkarni#2, Shruti Gadkari#3, Soumya Krothapalli#4

# Department of Information Technology, P.E.S. Modern College of Engineering,University of Pune
Maharashtra,India.

1abhijitbhosale67@gmail.com
2kulkarniabhishek20@gmail.com
3shruti.gadkari1994@gmail.com
4kmsoumya09@gmail.com

***Abstract*-Opinions or sentiments are of vital importance to all human activities and influence our behavior as well as decision making process. With the explosive growth of social media (eg. Reviews, forums, Facebook, Twitter, micro-blogs and posts in social networking sites) on the Web; the ease with which individuals or organizations can express their opinions has increased. Since the volume of this data is huge, manual analysis of these expressed opinions is a time consuming task. Thus, automated sentiment analysis system is the need of the hour. This paper presents an overview on Sentiment Analysis or Opinion Mining. It also enlists the research challenges of Sentiment Analysis.**

***Keywords*-Natural Language Processing (NLP),Hadoop Distributed File System,Named Entity Extraction,Information Extraction,Machine Learning**

I. INTRODUCTION

With the explosion of World Wide Web in the early 21st century the internet has touched every facet of our life.Further with the invent of the various social networking sites like Twitter, Facebook, micro-blogs etc , our current society has become free to express its opinions on this networking sites. Sentiments or opinion are so important that whenever we need to make any decision we hear other's opinion.One revolutionary aspect of social media is that they provide an outlet for instant,feedback from the public.Most of the sentiment or opinions expressed on these sites can give key insides for any topic provided.So, it is important to analyse the sentiments to extract positive, negative or even neutral opinions. Thus sentiment expressed over the internet by the people has gained the vital importance.

Sentiment analysis shows you the sentiment of an author while typing a social media message. In business sentiment analysis is becoming increasingly important as organisations want to know what their customers are thinking about their product. The accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgements. So, sentiment analysis is important for any organization; as without a clear understanding of sentiment surrounding their product ,their analytics alone can be misleading. This paper throws light on various approaches for carrying out sentiment analysis, some of the technologies that can be used and applications of sentiment analysis. In the end, the paper describes some of the complex tasks involved in the sentiment analysis.

II. SENTIMENT ANALYSIS (SA) AND OPINION MINING (OM)

Sentiment analysis or Opinion Mining is a computational study of people's opinions, attitudes and emotions towards an entity. Emotions are feelings generated from both conscious and unconscious processing. The emotional assessment of a situation is a general evaluation of a situation that affects our opinions and decisions that we make [1].Sentiment analysis refers to a broad area of natural language processing, computational linguistics and text mining to identify and extract subjective information from the available data. The main aim is to determine the attitude of the speaker; which may be their judgement or evaluation, their affective state or the intended emotional communication. The two expressions SA or OM are interchangeable and express a mutual meaning. However, some researchers stated that OM and SA have slightly different notions. Opinion Mining extracts and analyses people's opinion about an entity while Sentiment Analysis identifies the sentiment expressed in a text and then analyses it[2]. Sentiment Analysis involves Natural Language Processing and Information Extraction task which aims to obtain author's sentiments expressed either positively or negatively by processing and analysing numbers of documents.Essentially, sentiment analysis aims to determine the tone or attitude of the author with respective some topic in a document.[3]

Bing Liu et al. (2009) defines a sentiment or opinion as a quintuple-
{O_j, F_{jk}, SO_{ijkl}, H_i, T_l }
where
O_j is a target object,
F_{jk} is a feature of the object Oj,
H_i is an opinion holder,
T_l is the time when the opinion is expressed,
SO_{ijkl} is the sentiment value of the opinion of the opinion holder Hi on feature F_{jk} of object O_j at time T_l; SO_{ijkl} is +ve,-ve, or neutral, or a more granular rating.[4]

III. DIFFERENT LEVELS OF SENTIMENT ANALYSIS

A.Document Level:

The basic information unit is a single document of opinionated text. Sentiment analysis at the document level provides an overall opinion on an entity, topic or event which is expressed in a document. Here we aim to find out a whole document expresses a positive or negative sentiment. For example, a small report on a political event just before an election which expresses positive or negative an opinion about that. In this single review about single topic is considered. Thus, it is not applicable to documents which evaluate or compare multiple entities/topics (such as forum discussions, blogs, and news articles). The challenge in document level is that, each sentence in the document may or may not be relevant to the topic. In many cases it is hard to determine whether a document actually evaluates the entities[4].

B.Sentence Level:

In the sentence level sentiment analysis, the polarity of each sentence is calculated. The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion.In case of simple sentence, sentence contains single opinion. But a sentence may contain a complex opinionated text; in such cases use of sentence level must be avoided. Conditional sentences and sarcastic sentences need to be handled in sentence level. Sentence level sentiment classification cannot deal with opinions in comparative sentences. The advantage of sentence level analysis lies in the subjectivity/objectivity classification.[5].

C.Aspect Level:

The disadvantage of document level and sentence level analysis is that they don't determine what exactly people want to say. to overcome this disadvantage aspect level performs detail analysis involving feature based opinion mining and summerization.It concentrates on opinion of authors instead of looking at various language constructs.this helps understand sentiment analysis better.For example,"*LED TV picture quality is good but its sound quality is too cheap*", tells that the sentence is not entirely positive. In fact the sentence is positive about picture quality of LED TV but negative about its sound quality. [5].

IV. SUBJECTIVITY AND OBJECTIVITY

Nowadays feature based opinion mining technique is being used on a large scale to identify product features and user's opinions over them. But the part of reviews may or may not contain useful opinion or comments about product. Therefore knowing whether an opinion being expressed or a fact is being presented forms an essential part for analysing a dataset.
Subjective sentences are relevant text or the text which express some sentiment about the product. In contrast, objective sentences are irrelevant texts or they represent factual information.
Consider the review sentence:
The look of the mobile is amazing.
Mobile is a good device for connecting people over the globe.
Both sentences contain some opinion about mobile, having adjectives such as good and amazing. as per the stated definition the first sentence is subjective in nature and second one is in objective in nature[6]. This type of classification can be performed at either document level or sentence level. In document level classification involves identifying the subjective sentences from large collection for further processing.
Customer reviews at merchant sites contain star rated reviews. Higher rated reviews or documents can be placed in subjectivity class whereas lower rated reviews or documents can be classified as objective class. But as per the study in[7]many reviews or documents contain combination of both subjective and objective sentences. For example a movie reviews can be considered in subjective class as the reviews represent features of the movie. But on the other hand there can be sentences representing a factual information about an actor. Similarly we can consider a newspaper article in objective class. But these articles may contain some subjective sentences. Hence many research efforts in [8]have stated that the good indicator for the subjective can be identified by analysing the adjectives in a particular sentences.However, classification of a sentence as either subjective or objective is a non-trivial task due to unavailability of training dataset.In fact, annotated sets of subjective and objective sentences are difficult to obtain and requires a lot of manual processing and thus time consuming[7].

V.SUPERVISED MACHINE LEARNING

Supervised Machine Learning refers to the task of inferring a function from labeled training data. The training data consists of training examples which are used to map new examples. This is done by producing an inferring function[9].Below we mention some of the supervised machine learning algorithms :

A.Naïve Bayes

Nave Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the Bag of Words feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(label/feature) = \frac{P(label) * P(feature/label)}{P(feature)}$$

P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features—label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred [10]. Given the Nave assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(label/feature) = \frac{P(label) * P(f1/label) * \ldots\ldots * P(fn/label)}{P(feature)}$$

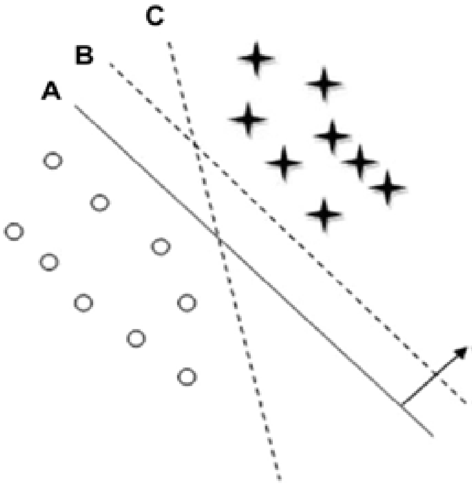

Figure 1: using support vector machine for classification problem[10].

B.SVM

SVM aims to maximize the margin around the separating hyperplane. In Figure1 there are 2 classes x, o and there are 3 hyperplanes A, B and C.

Hyperplane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation. Support vector machines (SVM),is considered the best text classification method (Rui Xia, 2011; Ziqiong, 2011; Songho tan, 2008 and Rudy Prabowo, 2009).SVM classification best suits for text data because of the sparse nature of text.These text contains few features that are irrelevant, but they are generally correlated with one another and organized into linearly separable categories[11]. Hyperplane are based on support vectors where,support vector is critical point close to decision boundary.

C.Decision Tree Classifiers

Decision tree is a classification scheme which represents a model of different classes which is used to generates tree & set of rules.Decision tree is a flow-chart-like tree structure, in which internal node denotes a test on an attribute and branch represents an outcome of the test. Also a node without children i.e leaf nodes represent class labels or class distribution.In decision tree acquired knowledge is represented in tree form, which is intuitive and easy to assimilate by humans.Decision tree classier is used to divide the data based on the condition on the attribute value. Main Algorithms includes Hunts algorithm,ID3,C4.5,CART,SLIQ,SPRINT[12].

VI. TECHNOLOGIES USED

Below we mention sone of the technologies that can be used for implementing Sentiment Analysis:

A.Hadoop:

Historically, unstructured data has been very difficult to analyze using traditional data warehousing technologies. New cost effective solutions, such as Hadoop are used. It uses the divide and rule methodology for processing. It is used to handle large and complex unstructured data which doesn't fit into tables. Hadoop Distributed File System (HDFS) is a distributed file system which runs on commodity machines. It is highly fault tolerant and is designed for low cost machines with high throughput. Posts can be downloaded and loaded into Hadoop using familiar tools like SQL Server Integration Services, or purpose built tools like Apache Flume. Once the data is loaded into Hadoop the next step is to transform it into a format that can be used for analysis and it is done using a process called Map Reduce. Map Reduce jobs can be written in a number of programming languages, including Java, Python,or can be generated by Hadoop Ecosystem tools such as Hive or PIG.The meaningful data now stored in Hadoop can be loaded into existing enterprise business intelligence (BI) platform or analysed directly using powerful self-service tools[13].

B.SAP HANA:
SAP HANA is an in-memory, column-oriented, relational database management system . HANA is designed to handle both large number of transactions and do complex query processing on the same platform.It is used for accessing, analysing, and interpreting massive volumes of unstructured data.SAP HANA is used for predictive analytics by discovering new interrelationships among the data thus revealing some new opportunities and hidden risks. With SAP HANA platforms powerful, in-database analytics, we can rapidly process structured, text, and spatial data for unprecedented insight and sentiment analysis[14].

C.MongoDB:
MongoDB is a cross-platform document-oriented database.MongoDB being a NoSQL database, uses table-based relational database structure that uses JSON-like documents. It includes dynamic schemas, which helps in integration of data in certain types of applications easier and faster.The aggregation framework enables users to obtain the kind of results for which the SQL GROUP BY clause is used[15].

VII. APPLICATIONS

sentiment analysis or opinion mining has gain lost of importance in recent years in the decision making process of organizations.Numerous businesses use the techniques of sentiment analysis for customer interest tracking and market perception,etc.The reviews of customer about products and services helps organizations to improve their quality of service and thus obtain financial benefit. There are numerous news items, articles, blogs, and tweets about each public company, which can be used to combine the overall opinion about the companies[16]. Some of the applications of sentiment analysis include online advertising, hotspot detection in forums, product review analysis etc. With the growing use of social media, the need of analysing the data is increasing.
Sentiment analysis can be used for:

- Tracking collective user opinions of products and services
- Analysing consumer trends, competitors and market buzz
- Measuring response to company-related events and incidents
- Monitoring critical issues to prevent negative viral effects
- Evaluating feedback in multiple languages

Sentiment analysis can applied to review classification, review summarization, synonyms and antonyms extraction, opinions tracking in online discussions, etc.

VIII. COMPLEX TASKS IN SENTIMENT ANALYSIS

A.Named Entity Extraction:
In Sentiment Analysis, it is important to know what are the target objects, or who are the competing entities. In simple scenario where we know the objects or entities , Named Entity Extraction is easy. But if we are a Third Party and want to know what people are talking about , then it is very hard to to predict the target objects.This is also hard because for the same entity ,the web users may write about it in many different ways. For example 'Motorola' may be written as 'Moto' or 'Mot'. Thus Named Entity Extraction is required for matching the entities discovered in corpus to the entities provided by the use [4, 17].

B.Co-reference Resolution:
This refers to the task of determining the references of multiple expressions in a sentence or multiple documents to one. For Example a blogger, typically refers to the blog which he had written before.Also users on social media sites like Facebook comment referring to some other comments.Thus it is important to resolve the issue of co-reference for an achieving efficiency [4, 17].

C.Relation Extraction:
Users generally use relations like "My mother liked the phone" or "My sister says that this phone is too cheap" to express their sentiments. This relations between the users expressing sentiments makes sentiment analysis hard[4, 17].

IX. CONCLUSION

Sentiment analysis is a Machine Learning problem that has been widely studied and researched in recent years.Since sentiment analysis is a complex task which includes a number of sub tasks such as Named Entity Extraction , co-reference resolution , Relation extraction etc , a fully automated and efficient system has not been developed till today.This is also due to the large vocabulary and nature of natural language .Thus we are able to solve the challenges mentioned above, it will be a step towards development of an efficient and fully automated sentiment analysis system.

References

[1] Meena Rambocas ,Joo Gama , *M̈arketing Research: The role of Sentiment".*

[2] *Sentiment Analysis*(n.d) [Online] Available: http://en.wikipedia.org/wiki/Sentiment_analysis .

[3] Subhabrata Mukherjee , *S̈entiment Analysis A Literature Survey".*

[4] Bing Liu, *S̈entiment Analysis and Opinion Mining"* Morgan & Claypool Publishers, 2012.

[5] Dudhat Ankitkumar M, Prof. R. R. Badre, Prof. Mayura Kinikar *"A Survey on Sentiment Analysis and Opinion Mining"* in International Journal of Innovative Research in Computer and Communication Engineering Vol. 2, Issue 11,2014 pp. 6634-6635.

[6] Ahmad Kamal , *S̈ubjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources"*, New Delhi , India.

[7] E. Riloff and J. Wiebe, *L̈earning Extraction Patterns for Subjective Expressions"*, Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-03), 2003, pp. 105-112.

[8] R. Bruce and J. Wiebe, *R̈ecognizing Subjectivity:A Case Study of Manual Tagging"*, Natural Language Engineering, 5(2), 1999, pp. 187-205.

[9] *Supervised learning* (n.d.) [Online]. Available:http:http://en.wikipedia.org/wiki/ Supervised_learning.

[10] Walaa Medhata,Ahmed Hassanb, Hoda Korashyb, *S̈entiment analysis algorithms and applications: A survey"* ,1999, pp. 187-205.

[11] Joachims T, *P̈robabilistic analysis of the rocchio algorithm with TFIDF for text categorization"* , Presented at the ICML conference, 1997.

[12] Prof. Navneet Goyal *D̈ecision Tree Classification"* , [Online]. Available FTP: csis.bits-pilani.ac.in/faculty/.../Decision%20Tree%20 Classification.ppt File: Decision Tree Classification.ppt

[13] *Apache Hadoop* (n.d.) [Online]. Available:http://en.wikipedia.org/wiki/Apache Hadoop

[14] *SAP HANA* (n.d.) [Online]. Available:http://en.wikipedia.org/wiki/SAP_HANA

[15] *MongoDB* (n.d.) [Online]. Available:http://en.wikipedia.org/wiki/MongoDB.

[16] *Techniques and applications for sentiment analysis* (n.d.) [Online]. Available:http://www.ceine.cl/techniques-and-applications-for-sentiment-analysis/.

[17] International School of Engineering."CPEE (CSE 7206c) - Sentiment Analysis (Part 3)
- Building Sophisticated Sentiment Extraction System,"Youtube,Dec.9,2013[Video File].Available:
https://www.youtube.com/watch?v=Quvf3TLI4Js