

**Sentiment Analysis: A Review**Sneha Jadav¹, Bhavesh Tanawala², Hetal Guadani³¹Computer engineering, Birla Vishvakarma Mahavidyalaya, Vallabh vidyanagar-388120²Computer engineering, Birla Vishvakarma Mahavidyalaya, Vallabh vidyanagar-388120³Computer engineering, G.H.Patel College of Engineering and Technology, Vallabh vidyanagar-388120

Abstract — Sentiment Analysis is one of the applications of Natural Language Processing. Due to the ample amount of business and social applications people are anticipated to build up a system which can discover sentiment lies in the electronic text. This paper presents different approaches to analyze persons' opinions and challenges that emerge in this field.

Keywords- sentiment, opinion, natural language processing (NLP), lexicon, machine learning

I. INTRODUCTION

Sentiment analysis could be one form of natural language processing which is used to know the individuals' sentiment regarding a specific product or topic or service. Sentiment analysis, that is additionally referred to as opinion mining, requires to construct a system to gather and examine opinions about the product or service described in blog posts, comments, reviews or tweets.

Sentiment analysis can be utilized in alternative ways. Like, for an online shopping website it eases in deciding whether or not products are likeable or unlikeable by individuals. Using this knowledge one can improve marketing. And also in social media one can analyze which topic is currently trending and whether people are talking positive or negative about it. Sometimes this type of information leads to some conclusion or prediction i.e. polls, elections etc.

As it has many advantages, various challenges are also there in Sentiment analysis i.e. some words contain positive sentiment at one time and at another time it can be counted as negative sentiment. For example, "This phone is good" is not similar from "This phone is not good" as both are showing different sentiment lies in the sentence. Another problem is that every individual has different ways to express their views which can't lead to actual final conclusion. Many opinions can have both positive and negative sentences within them. In social medium like twitter or blogs, people can write different sentiments within one sentence or one review that is easy to understand for human but for computer system it's difficult to find true sentiment. Sometimes even human can't understand what someone is trying to say based on small part of text because it doesn't contain actual context. For example, "That updated version of application was as good as its last version" is entirely based on the sentiment of person regarding last version. An individual's decision is mostly dependent upon other's reviews about that product or service. So this area gives motivation to build such systems which can process these reviews and recommend that product to other users.

Opinion mining majorly works on people's sentiments and analysis can be done by traditional text mining techniques. Some areas of research are there in sentiment analysis like sentiment classification, Sentiment classification related to features and opinion summarization. To classify the whole document based on reviews, sentiment classification is used. However, to make classification based on features of particular objects, feature-based Sentiment classification is used. Opinion summarization mines product features only on which other individuals have made reviews. So opinion summarization is completely distinct from conventional text summarization because standard text summarization includes summarization of reviews by selecting some parts or rewriting some of the sentences from the reviews to extract main parts and opinion summarization doesn't.

Till now English and Chinese languages are majorly used for sentiment classification but on remaining languages, less research had been done. In this study, more work is done for English language and some for Chinese. Sentiment analysis firstly introduced in late 1990's but at that time it was not well known. From 2000, it has become area of interest for researchers as the use of online websites and social mediums is increased.

To do sentiment analysis Data set is necessary to classify the text or to process the text. Nowadays there is huge amount of web content so we can easily work on large set of data which makes classification more accurate. These sources can be review sites, micro-blogs, tweets, social sites etc. Some online sites provide API i.e. Twitter to make use of their data while for others we can use some other techniques like web crawling, web scraping etc. Some online libraries also provide data sets.

Now after getting dataset sometimes pre processing of data is required before initiating classification methods on datasets. There are various pre processing techniques available nowadays which includes stop word removal, tokenization, stemming etc. Applying these techniques on data set makes data easy to classify because it stores only those words or sentences which contain useful information that here is sentiment.

This paper is organized as follows: Section II describes/ portrays work done so far on classification based on both approaches that one is lexicon based and other is machine learning. Then various tools that are available for sentiment classification are presented in section II only. Section III summarizes our study and presents some future work for research.

II. METHODS

To extract the polarity of sentiment is the main issue in opinion mining for which some approaches are already introduced in past with different methods.

Basically three types of approaches are present for sentiment analysis. First one is lexicon based (Dictionary Based) and another one is machine learning approach. Here machine learning approach again contains two different techniques one is supervised machine learning and other is unsupervised machine learning approaches.

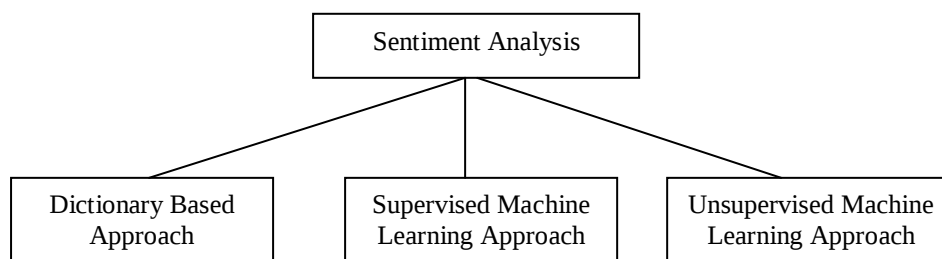


Figure 1. Types of methods

Also hybrid methods are there for classification problems which mix two or more above techniques.

III. DICTIONARY BASED APPROACH

In lexicon based(Dictionary based) approach, dictionary is created for positive and negative words and for every word dictionary is checked to see whether word is belonging to positive or negative words' dictionary. Some approaches also included weight for each word in dictionary according to that overall sentiment score is calculated. SentiWordNet, WordNet, ConceptNet etc are mostly used lexicon resources. In these all resources SenticNet is widely used. SentiWordNet has different versions i.e. SentiWordNet 1.0, 2.0 and now 3.0 is also available.

In paper [8], SENTIWORDNET 3.0 is developed by authors which is a lexical resource mainly invented for applications of sentiment analysis and opinion mining. It is an advanced version of SENTIWORDNET 1.0 which can be easily used for analysis purpose. Each SENTIWORDNET 1.0 and 3.0 are the outcome of automatically annotating all WORDNET synsets in line with their degrees of positivism, negativity, and neutrality. SENTIWORDNET 1.0 and 3.0 take issue (a) within the versions of WORDNET which they annotate (WORDNET 2.0 and 3.0, respectively), (b) within the algorithm used for automatically annotating WORDNET, which currently includes (additionally to the previous semi-supervised learning step) a random-walk step for refining the scores. We tend to here discuss SENTIWORDNET 3.0, particularly immersion on the enhancements regarding facet (b) that it embodies with regard to version 1.0. Final results indicate accuracy enhancements of about 20% with regard to SENTIWORDNET 1.0.

In paper [9], authors had found accuracy of classification is matter of concern because sentiment analysis techniques which are developed before couldn't classify data accurately and produce biased results which are not good. Use of only lexical content is not useful. So, they have applied sentiment analysis using lexicon or dictionary based approach which make use of word sense information and make advanced technique to classify the data correctly. Authors have made this unique technique using lexical resources like SentiWordNet and WordNet as well as Word Sense Disambiguation. So they applied the SentiWordNet lexical resource and Word Sense Disambiguation for finding political sentiment from real time tweets. They also introduced technique to handle negation as one of pre-processing steps to increase the accuracy.

[10]During this paper, unique sentiment analysis model is invented by authors. This model has supported common-sense information which is extracted from ConceptNet based ontology and context info. ConceptNet based ontology is employed to seek out the domain specific ideas that successively fashioned the domain specific vital features. Then, the

polarities of the extracted ideas are determined using the contextual polarity lexicon that they tend to developed by considering the context info of a word. Finally, linguistics orientations of domain specific features of the review document are aggregated based on the importance of a feature with related to the domain. The importance of the feature is achieved by the depth of the feature in the ontology. Experimental results show the effectiveness of the projected strategies.

[11] This paper given other lexicon-based sentiment analysis algorithmic rule that has been designed with the most specialize on real time Twitter content analysis. The algorithm consists of two key parts, specifically sentiment normalization and evidence-based combination function that are employed in order to estimate the intensity of the sentiment rather than positive/negative label and to support the mixed sentiment classification method. Finally, they illustrated a case study examining the relation between negative sentiments of twitter posts associated with English Defense League and also the level of disorder throughout the organization's connected events.

In paper [12], authors have done analysis of public info from social media which may yield attention-grabbing results and insights into the globe of public opinions regarding nearly any product, service or personality. Social network data is one in every of the foremost useful and precise indicators of public sentiment. In this paper they have mentioned a technique that permits utilization and interpretation of twitter data to see public opinions and analysis was done on tweets regarding the iPhone 6. They conjointly enclosed feature specific popularities and male-female specific analysis. Mixed opinions were found however general consistency with outside reviews and comments was determined.

IV. SUPERVISED MACHINE LEARNING APPROACH

This approach is known as supervised learning because datasets used for it are labeled. First of all classifier is trained according to those labels and after that testing phase is carried out for unlabelled datasets. For this approach, various classification algorithms are present i.e. Naive Bayes, SVM (Support Vector Machine), random forest etc.

Authors Xing Fang, Justin Zhan [2] did experiments for two types of classification. First, sentence-level classification and review-level classification and also achieved good results. For this study, reviews of products from online website www.amazon.com Data were used. And some supervised machine learning algorithms were used for classification like Naive Bayesian, SVM, Random Forest etc. However some limitations are also there. First one is that it is not easy to classify reviews if reviews are having star-scaled ratings and second is that if review contains implicit sentiment then it can't be detected by this invented technique as it is based on frequency of words.

Authors Pragya Tripathi, Santosh Kr Vishwakarma, Ajay Lala used Rapid Miner tool[4], that facilitates in building the classifier additionally as ready to use it on the testing dataset. Tweets from twitter are used as dataset. They defined number of the future work like (1) they goal to increase the size of the testing dataset and search for the complicated troubles associated with it. (2)The number of labels can also be elevated or modified according to requirements (3) they purpose to use emoticons additionally for classification in future.

As a replacement method D V Nagarjuna Devi, Chinta Kishore Kumar, Siriki Prasad[5] have developed Sentiment analysis approach which helps in finding which feature is better so it is known as feature based sentiment analysis. They have used a classifier called Support Vector Machine (SVM) in a novel approach. This approach is highly effective in ways like if we want to find the users' sentiment regarding any product's specific feature or about the whole product. They have defined to work on sarcastic expressions as a future work as these expressions are hard to understand by people and also by the computer system. Spam detection is also challenging problem from reviews of people. To get the solution of co-reference resolution, this study can be extended. Some authors conjointly highlighted comparison between varied classification techniques so we are able to decide which one is good for that sort of information.

Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt[6] have performed similar work that shows a comparison between the results obtained by exploiting the following machine learning algorithms: Naïve Bayes and Support Vector Machine and an Artificial Neural Network model: Multilayer Perceptron. They defined future work as to aim to classify using smiley or emoticons.

In paper [7] authors also introduced new feature selection method on large amount of movie review data set which is Gini Index for better accuracy and it has reduced error rate also. They defined implicit sentiment as future work.

Authors	Dataset	Method	Accuracy	
Xing Fang, Justin Zhan(2015)[2]	Product review data	Supervised machine learning	F1-Avg scores	Random Forest=0.85
				Naive Bayesian=0.6
				SVM=0.91
Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang(2016)[3]	Micro blogs	Hybrid method	80%	
Pragya Tripathi, Santosh Kr Vishwakarma(2015)[4]	Tweets	Supervised machine learning	Naive Bayesian classifier= 63.33%	
			K-NN classifier= 70%	
D V Nagarjuna Devi, Chinta Kishore Kumar,Siriki Prasad(2016)[5]	Product reviews	Supervised machine learning	For SVM = 88.13%	
Nehal Mamgain, Ekta Mehta, Ankush Mittal,Gaurav Bhatt(2016)[6]	Tweets from AIIMS,IITs and NITs	Supervised machine learning	AIIMS	Naive Bayes=90%
				SVM=91%
				Neural Network=92.6%
			IITs	Naive Bayes=90%
				SVM=83%
				Neural Network=89.6%
			NITs	Naive Bayes=83%
				SVM=81%
				Neural Network=87.6%
Asha S Manek, P Deepa Shenoy,M Chandra Mohan, Venugopal K R(2016)[7]	Movie reviews	Supervised machine learning	SVM(Linear)	Small dataset=97.25%
				Large dataset=94.46%
R. Jose and V. S. Chooralil(2015)[9]	Tweets	Dictionary based approach	78.6%	
Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg(2015)[10]	Restaurant, software and movie reviews	Dictionary based approach	Restaurant reviews=65.7%	
			Movie reviews=70.1%	
			Software reviews=67.8%	
Anna Jurek, Maurice D. Mulvenna and Yaxin Bi (2015)[11]	Stanford test Twitter corpus and IMDB movie reviews	Dictionary based approach	Tweets = 70.6%	
			Movie reviews=64.82%	
Syed Akib Anwar Hridoy, M. Tahmid Ekram, Mohammad Samiul Islam, Faysal Ahmed and Rashedur M. Rahman (2015)[12]	Tweets	Dictionary based approach	96%	
Andrius Mudinas, Dell Zhang, Mark Levene(2012)[13]	CNET software reviews and IMDB movie reviews	Hybrid method	Software reviews=81.57%	
A.-M. Popescu and O. Etzioni(2005)[14]	Product reviews	Unsupervised machine learning	Opinion extraction(Precision and Recall):+0.08 & -0.02 Opinion Polarity(Precision and Recall):+0.06 & -0.04	

Table 1 Literature Survey

III. UNSUPERVISED MACHINE LEARNING APPROACH

Unsupervised learning is where we only have input dataset and it is unlabelled. As all data is unlabeled, the algorithms learn to inherent structure from the input data. Unsupervised machine learning includes clustering and association. Less work is done using this approach as it is not as accurate as supervised machine learning.

An unsupervised data mining system is invented by authors in paper [14] called as OPINE to extract the reviews to build a system which contains features of products, their evaluation by individuals (reviewers), and their relative quality across products. Relaxation labelling of OPINE is used for finding the semantic orientation of tokens (words) in context results in increase in performance on the tasks of searching sentiment phrases and their polarity.

IV. HYBRID APPROACH

Nowadays, combined approaches are also used by many authors in which two or above approaches are combined to achieve better accuracy. There are two types of hybrid methods: Linear hybrid approach and Parallel hybrid approach[15]. Linear hybrid approach is defined as one technique's output is given as input to another method while in parallel hybrid technique both approaches are mixed for classification.

In paper[3], micro blogs were used as dataset. Authors Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang had invented hybrid method to classify multi-class categorization which mixes the model based technique with the lexicon-based technique. As features they have used emoticons and Naïve-Bayes classification. But it considered less features for Naive-Bayes classification building and it labeled less test data set. Better accuracy was also observed.

In paper [13], authors have developed the anatomy of P-Senti model which is used to analyze the data based on concepts and perfectly combines dictionary-based sentiment analysis and machine learning techniques. They have used IMDB movie reviews as training dataset.

V. CONCLUSION

The study solves the primary problem of sentiment polarity classification and sentiment summarization. This paper mainly focuses on different techniques which were already proposed with their benefits and whether any future work is possible or not.

REFERENCES

- [1] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
- [2] Xing Fang, Justin Zhan, "Sentiment analysis using product review data", Journal of Big Data, 2015.
- [3] Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang, "A hybrid method for multiclass sentiment analysis of micro-blogs," 13th International Conference on Service Systems and Service Management (ICSSSM), Kunming, 2016, pp. 1-6.
- [4] Pragya Tripathi, Santosh Kr Vishwakarma, Ajay Lala, "Sentiment Analysis of English Tweets Using RapidMiner", International Conference on Computational Intelligence and Communication Networks, 2015.
- [5] D V Nagarjuna Devi, Chinta Kishore Kumar, Siriki Prasad, "A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine", IEEE 6th International Conference on Advanced Computing, 2016.
- [6] Nehal Mangain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt, "Sentiment Analysis of Top Colleges in India Using Twitter Data", International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016.
- [7] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, Venugopal K R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", et al. World Wide Web, 2016.
- [8] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.
- [9] R. Jose and V. S. Choorailil, "Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation," 2015 International Conference on Control Communication & Computing India (ICCC), Trivandrum, 2015, pp. 638-641.
- [10] Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg, "Sentiment Analysis Using Common-Sense and Context Information," Computational Intelligence and Neuroscience, vol. 2015, Article ID 715730, 9 pages, 2015. doi:10.1155/2015/715730.
- [11] Jurek, A., Mulvenna, M.D. & Bi, Y. Secur Inform (2015) 4: 9. doi:10.1186/s13388-015-0024-x
- [12] Anwar Hridoy, S.A., Ekram, M., Islam, M.S. et al. Decis. Anal. (2015) 2: 8. doi:10.1186/s40165-015-0016-4

- [13] Andrius Mudinas, Dell Zhang, Mark Levene, "Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis", WISDOM'12, August 12, 2012, Beijing, China.
- [14] A.-M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," in proceedings of the Conference on Human Language Technology and Conference on Empirical Methods in Natural Language Processing 2005 (HLT/EMNLP 2005). ACL, 2005, pp. 39–346.
- [15] K. Schouten and F. Frasincar, "Survey on Aspect-Level Sentiment Analysis," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813-830, March 1 2016.