

A Study on Sentiment Analysis Algorithms

Lohita Sivaprakasam

SSN College of Engineering, Chennai

Aarthe Jayaprakash

SSN College of Engineering, Chennai

ABSTRACT:

Sentiment Analysis refers to the systematic study of the range of emotions portrayed in any form of text and the categorization of the expressions studied as positive, negative or neutral. This paper aims at presenting ideas inculcated in existing literature for sentiment analysis in four major domains where the analysis can be efficiently utilized. This study clearly reveals the need for novel techniques that can improve the current status of research in the area of sentiment analysis.

KEYWORDS:

Sentiment Analysis, Classifiers, Support Vector Machine, Naive Bayesian

INTRODUCTION:

Sentiment analysis is mainly performed on the datasets which are predominantly the opinions voiced by the consumers. Exacting the attitude of the consumers towards certain applications helps in providing recommendations customized to an individual such that it fits his/her interests. The decision making process for the vendors is enhanced by means of sentiment analysis. The overall contextual polarity to any event or document is determined, which paves the way for knowing the pulse of the people as a reaction to the analyzed subject or topic.

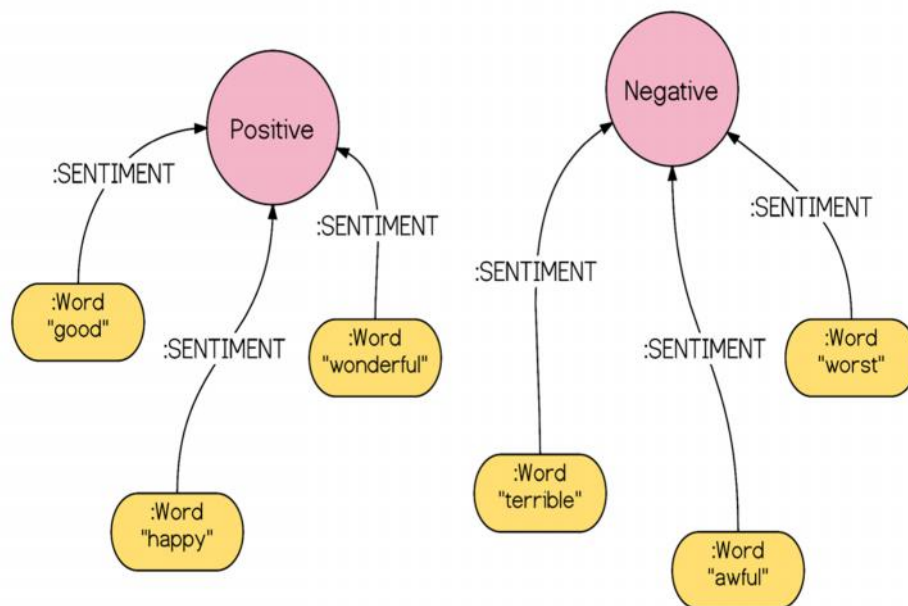


Fig 1: Categorization of sentiments based on polarity

Given a piece of text that reads:

“I love the summer in Norway, but I hate the winter.”

A weak sentiment analysis system categorizes “love the summer” as a positive sentiment and “hate the winter” as a negative sentiment, but reports the entire piece of text as effectively neutral. This is done by cancelling the positive “love” and the negative “hate”.

A strong sentiment analysis system categorizes text based on proper classification of polarity as well as based on subjectivity or objectivity of the writer of the text.

GENERAL FLOW OF SENTIMENT ANALYSIS:

Part 1- Identifying sentiment in phrases:

From a given piece of text, patterns are identified and useful phrases containing sentiments are extracted. These sentiments are then analysed by an algorithm and are evaluated by scores from a sentiment dictionary. Another algorithm is applied on the same for intensification (i.e to find the intensity of a sentiment, since people generally have mixed emotions). Finally a machine learning algorithm is applied to see if the syntax of the text affects its sentiment. The output of part 1 is a collection of weighted sentiment phrases.

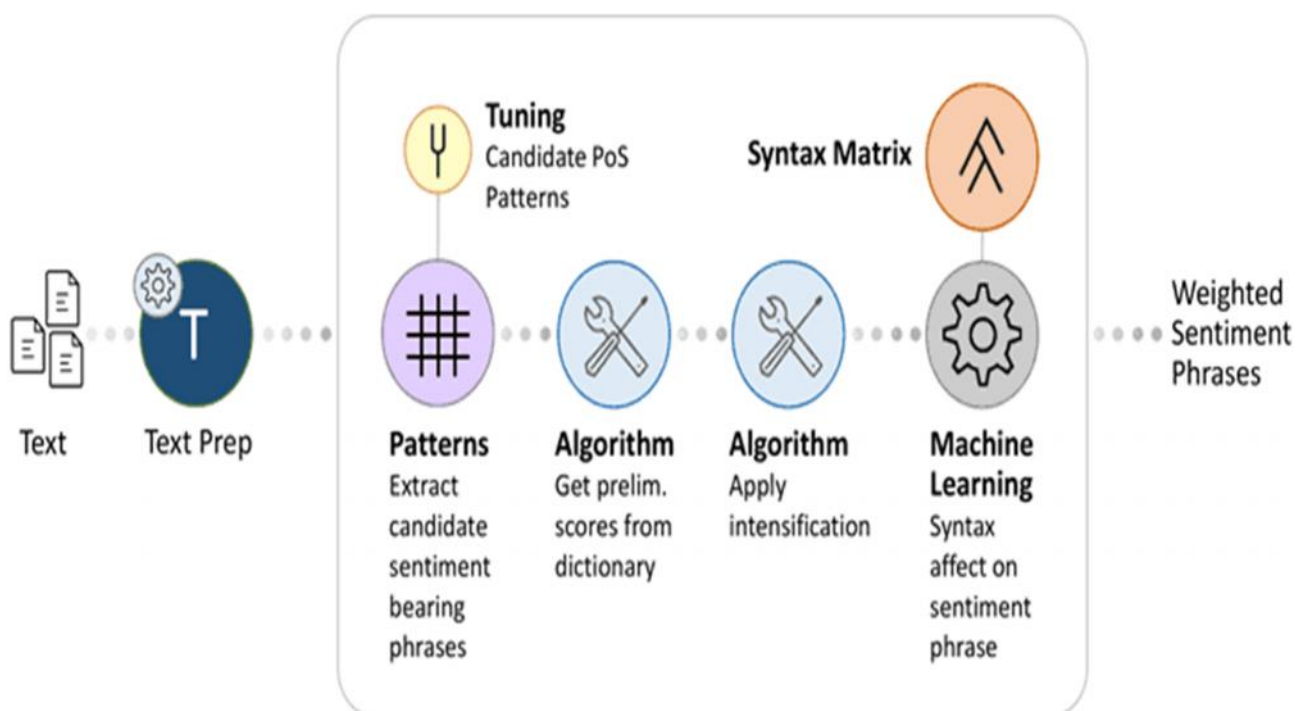


Fig 2: Generation of weighted sentiment phrases

Part 2- Identifying sentiment in a document:

Part 2 categorizes text based on subjectivity and objectivity of the writer. Some people express views clouded by emotion, so they may exaggerate what is intended or not say enough. An algorithm can be used to categorize how accurate a piece of text is and evaluating a score based on this accuracy. This is added to the score evaluated in Part 1, for each phrase. The scores of all the phrases containing sentiments in the document are added up and the document’s sentiment is appropriately identified.

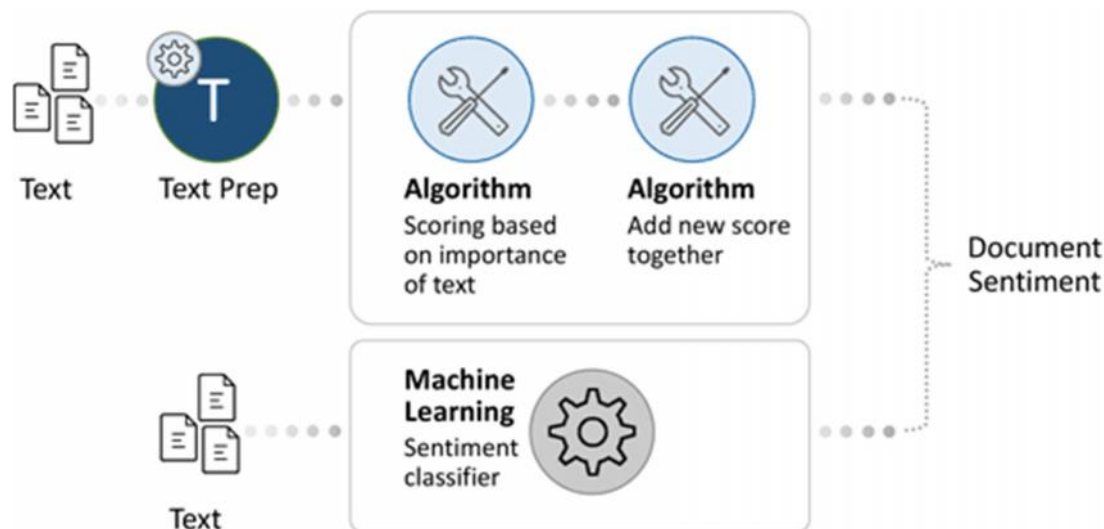


Fig 3: Identification of document sentiment

DOMAIN-SPECIFIC ANALYSIS:

Sentiment Analysis has a variety of applications across numerous domains.

In this review we are elaborating the use of the same across 4 major domains:

1. Product Reviews
2. Movie Reviews
3. Book Genre-based Classification
4. Sentiment analysis in news articles and social media

1. Product Reviews

There is a requirement for sentiment analysis of Product reviews to help both the customer and the vendor. Sentiment Analysis helps the customer choose the right product for purchase, given the specifications. It also helps the customer report his/her feedback to the vendor. The vendor can make use of the analysis of reviews to better their products and have an edge over their competitors from a business perspective.

Jian jian and Ping ji et. al., [1] proposed a methodology for extracting product feature specific opinions from customer reviews posted on the internet. They proposed a co-clustering algorithm to provide an outline of the customers' attitude towards the various perspectives of product specifications. Their method utilized the concept of conditional random fields that blends the various aspects of product specifications with the detailed reasoning behind the opinions provided by customers.

Subhabrata mukherjee and pushpak Bhattacharyya et. al., [2] deployed a dependency parsing methodology in order to map the various features with their corresponding opinions. This process comprises of two main steps- extracting relevant features from the reviews and clustering the opinions pertaining to each of these features. The experiments were performed using two datasets – dataset 1 [3] with an average accuracy of 80.98% and dataset 2 [4] with an average accuracy of 70%.

The techniques followed for the sentiment analysis in the field of product reviews can be deduced to three major ones, namely Support vector machine, Naive Bayesian and Maximum Entropy. From the general trends in the degree of accuracy obtained for each of the three mentioned techniques when performed on the product datasets[5], it can be observed that the degree of accuracy is highest for Support Vector Machine, followed by moderate degrees of accuracy for both Naive Bayesian as well as Maximum Entropy technique.

2.Movie Reviews

The sentiment analysis in movie reviews is required to know whether the opinion showcases a negative or positive feedback rather than performing a diligently accurate analysis.

M.Hiu and B.Liu et. al.,[6] follow a lexicon-based approach for sentiment analysis. The approach makes use of a list of keywords categorized as positive and negative, each word provided with a corresponding strength factor depicting the severity of the word. However,a shortcoming of this approach is that the opinion words are context-specific and hence pose a problem of the coverage area based on the lexicon size [7].

In order to recognize the sentiments based on domain-specific opinion keywords, B.Pang, L. Lee, and S. Vaithyanathan et. al.,[8] employed the machine learning approach. The learning phase is performed by taking into account the keywords, from the training datasets that belong to a particular domain. The setback is that machine learning is not time-efficient when the dataset is made scalable. Also,the analysis impudes a highly specific nature to the domain on which the training datasets were learnt[7].

Jenq-Haur WANG and Ting-Wei LIU et. al.,[9] formulated an approach that comprises of lexicon adjustment based on heuristic rules and opinion rating aggregation in order to provide movie recommendations with respect to the reviews fed on movies. The heuristic rules for lexicon adjustment are firstly to locate and take away the dessicated words that do not qualitatively evaluate the movie in entirety. Secondly, the opinion keywords included are chosen to be the trending ones in order to expand the coverage.

3.Book Genre-based Classification

The array of sentiments expressed in a story book can be extracted to categorize the book as belonging to a specific genre based on the sentiments analyzed. This categorization can be of help to provide recommendation for consumers based on the genre of books picked earlier by them.

There has been numerous research works into this field of study. K. Shimizu and M. Hagiwara et.al.,[10] utilizes the adjectives extracted from the text to propose estimation method of word image. They enhanced this approach by using a conflate of noun and verb to propose an impression estimation method of word image [11].

Nakamura et. al.,[12] developed an 'Emotion Expression Dictionary' which can be viewed as a mapping between emotion types and their corresponding sentiment expressions. In the dictionary, a range of 2171 sentiment expressions are categorized into the 10 primitive emotion types recognized by Nakamura. This Emotion Expression Dictionary is used to identify the sentiment genre of a story book.

4.Sentiment analysis in news articles and social media

The news articles published are subject to sentiment analysis to obtain the exact tone with which the piece of news is presented. Prashanth Raina et. al.,[13] proposed a methodology for sentiment analysis in news articles which is based on an opinion mining engine. The common-sense knowledge derived from Concept Net and Semantic Net are used in this methodology. Huge datasets ie. news article with quite a many sentences was used to test the accuracy of the proposed opinion mining engine. The testing resulted in a 71% accuracy in the classification while the precision of neutral sentences was as high as 91%.

The social media is a huge repository of people's opinions on almost all the issues faced by them. This huge range of opinions can be effectively utilized to understand the pulse of the people regarding a variety of issues. Federico Neri et. al.[14], did a study on over 1000 posts on Facebook to analyze the sentiment attached to the posts.

Ana CES. Lima et al.[15] developed an automated sentiment classifier to analyze the sentiment attached to the tweets posted by the people. The Naive Bayesian technique was used to develop the classifier. The major drawback is the strict classification of words into positive or negative with no room for neutral words. Min Wang et.al.[16] developed an approach to enhance the classification of sentiment expressions by computing a

quantitative value corresponding to each sentiment expression. The experiments carried out to check the effectiveness of this method were positive enough to prove the correctness and efficiency of this methodology.

COMPARISON OF TECHNIQUES:

The techniques used by various researchers in the field of Sentiment Analysis are reviewed and tabulated in Table 1.

Table 1. Distribution of techniques applied and accuracy achieved

Serial No.	Technique	Accuracy achieved	Year of publishing of paper
1.	Emoticon-base technique	89.56%	2012 [15]
	Word based technique	89.92%	
	Hybrid technique	91.94%	
2.	Support Vector Machine	81.43%	2013 [17]
3.	Dependency Parsing Using dataset 1 [3]	80.98%	2014 [2]
	Using dataset 2 [4]	70%	
4.	Semantic Analysis-WordNet	89.9%	2014 [18]
5.	Naive Bayesian approach	88.8%	2015 [19]

CONCLUSION:

This paper presents the concept of sentiment analysis in specific domains and the need for the same. It provided a review of the research works published in papers pertaining to each specific domain. In doing so, the aspects achieved so far are elucidated and the shortcomings of certain aspects have also been mentioned. It can be concluded that there is scope for further research and enhancements in the sentiment analysis of the specified domains.

REFERENCES:

- [1]Jian Jin and Ping Ji,"Mining online product reviews to identify consumers FineGrinedConcerns",IET,2015,Hong Kong SAR,China.
- [2]Subhabrata Mukherjee,Pushpak Bhattacharyya,"Feature Specific Sentiment Analysis for product Reviews",IET,2015,IIT Bombay.
- [3]Himabindu Lakkaraju, Chiranjib Bhattacharyya,Indrajit Bhattacharyya and Srujana Merugu,"Exploiting Coherence for the simultaneous discovery of latent facts and associated sentiments",SIAM International Conference on Data Mining (SDM),April2011.
- [4] Minqing Hu and Bing Liu,"Mining and Summarizing customer reviews", KDD 04: proceedings of the tenth ACM

-
- SIGKDD international Conference on knowledge discovery and data mining.
- [5] R. RajKumar,V.P. Kallimani, Lam H. Lee, Dino Isa, "Text document processing with naïve bayes and support vector machine", published by IEEE computer society.
 - [6] M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, 2004, pp.168-177.
 - [7] B. Liu, *Sentiment Analysis and Opinion Mining*, 2012, Morgan & Claypool Publishers
 - [8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumps up? sentiment classification using machine learning techniques," *Proceedings of the ACL conference on Empirical methods in natural language processing (EMNLP 2002)*, 2002, pp. 79-86.
 - [9] Jenq-Haur WANG and Ting-Wei LIU 'Improving Sentiment Rating of Movie Review Comments for Recommendation',2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), *National Taipei University of Technology, Taiwan*
 - [10] K. Shimizu and M. Hagiwara: "Image Estimation of Words Based on Adjective Co-occurrences," *IEICE TRANSACTIONS on Information and Systems*, J89-D, pp.2483-2490, 2006.
 - [11] K. Shimizu and M. Hagiwara: "A Method to Estimate the Impression of Noun-Verb Phrases," *Transactions of Japan Society of Kansei Engineering*, Japan society of Kansei Engineering, 10, pp.505-514, 2011.
 - [12] A. Nakamura: "Emotion Expression Dictionary", Tokyodo Publication, 1979.
 - [13] Prashant Raina., "Sentiment analysis in News articles using Senticomputing", *IEEE 131h International conference on Data Mining Workshops*, 2013, pp.959-962.
 - [14] Federico Neri, Carlo Aliprandi, Federico Capeci, Monsterrat Cuadros, Tomas, "Sentiment Analysis on social media", *IEEE/ACM International conference on Advances in Social Networks analysis and mining*,20 12,pp. 919-926.
 - [15] Ana C. E.S.Lima, Leandro N.de Castro., "Automatic Sentiment Analysis oftwitter Messages", *Publisher IEEE*, 2012, pp. 52-57.
 - [16] Min Wang, Hanxio Shi., "Research on Sentiment Analysis Technology and Polarity Computation of Sentiment words", *Publisher IEEE*, 2010, pp.331-334
 - [17] Warih Maharani, "Microblogging Sentiment Analysis with LexicalBased and Machine Learning Approaches", 2013 *IEEE*, pp. 439-443.
 - [18] Geetika Gautam et al., "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", *Publisher IEEE*,2014.
 - [19] Alexandra Cernian et al., "Sentiment analysis from product reviews using SentiWordNt as lexical resource", *Publisher IEEE*, 2015.