International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I4P107

Volume 69 Issue 4, 34-38, April, 2021
© 2021 Seventh Sense Research Group®

Review Article

A Survey on Social Digital Data-Based Sentiment Mining Techniques and Feature

Rajesh Sisodiya[1], Praveen Kumar Mannepalli[2]
[1]*Research Scholar, School of Computer Science & Technology , LNCT University, Bhopal*
[2]*Associate Professor, School of Computer Science & Technology , LNCT University, Bhopal*

Received Date: 07 March 2021
Revised Date: 18 April 2021
Accepted Date: 20 April 2021

Abstract - *Sentiment mining plays a very important role in business to understand the opinion of customers to improve the Customer of the products also depends on the opinion of others who have bought the products already. Reviews or feedback becomes the deciding factor for buy or sells a product. This paper has focused on elaborating the user rating behavior for a particular kind of service, product, news. Techniques developed by various researchers are discussed with their implementation dataset and outcomes. Some digital features are also detailed, which play an important role in increasing the accuracy for the prediction of sentiment class. Types of sentiment analysis and mining were also detailed. Paper has summarized evaluation parameter values of the sentiment data analysis for comparing techniques*

Keywords - *Data mining, Online Social content, Sentiment analysis, Text Clustering.*

I. INTRODUCTION

Our world has changed drastically in the last few decades. A person's sentiments are no longer shared only with his or her immediate family and friends but instead have the capability of influencing the decisions of thousands or millions of people the individual has never even met. The Internet has given the platform to broadcast grievances and recommendations that can reach across the world to an individual. The existence of public networks gives these sentiments the potential to snowball into a viral frenzy that can make your organization's products or services a worldwide boon or a whole catastrophe in just a matter of days [1].

The Sentiment Analysis is the computational investigation of individuals' feelings, frames of mind, and feelings toward a substance. The element can speak about any people, occasions, or themes. These subjects are destined to be secured by surveys. This recognizes the feeling communicated in a book at that point investigates it. Consequently, the objective of the feeling examination is to discover sentiments, distinguish the notions they express, and afterward arrange their extremity as they appeared [2]. Sentiment Analysis can be viewed as an arrangement procedure. There are three primary characterization levels in Sentiment Analysis: document level, sentence level, and aspect level. Document-level plans to characterize a conclusion archive as communicating a positive or negative supposition or estimation while Sentence-level intends to arrange to feel clarified in each sentence. The initial step is to distinguish whether the sentence is abstract or goal. On the off chance that the sentence is emotional, Sentence-level supposition examination will decide if the sentence communicates positive or negative sentiments [3]. Be that as it may, there is no essential contrast among document and sentence level orders since sentences are simply short archives [4]. Perspective level sentiment analysis means to group the sentiment as for the particular parts of substances. The initial step is to distinguish the substances and their viewpoints.

Finding sentiment sources and monitoring them on the Web can still be a difficult task because there are a large number of diverse sources, and every source may also have a big volume of sentimentated text (text with sentiments or sentiments). Selecting an attribute for sentiment classification using feature relation networks. In many cases, sentiments are hidden in long forum posts and blogs. It is complex for a human reader to find relational sources, extract relational sentences with sentiments, read them, understand them, and organize them into usable forms [5]. Thus, automated summarization systems are needed. Using this summarization, we can recognize the importance, quality, popularity of products and services. In this system, we make a summarization for the movie. But, we can use this system anywhere where text analysis is required. Sentiment analysis, also known as sentiment mining, grows out of this need. It is a challenging natural language processing or text mining problem. Due to its tremendous value for practical applications, there has been an explosive growth of both research in academia and applications in the industry

II. TYPES OF SENTIMENT MINING AND ANALYSIS

There are three types of level of sentiment mining [5, 6].

- Document-level Sentiment Mining: They classified documents according to sentiment instead of the topic. They contain in the form of forums or blogs, the

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

document level analysis is not desirable. So subjectivity/objectivity classification is used in this type of classification.

- Sentence level Sentiment Mining: In sentence-level Sentiment Mining, sentiment word is extracted from sentence-level and makes them as positive and negative.
- Feature level Sentiment Mining: The features that contain sentiment words are found out, and a phrase level classification is done. It is also known as feature-level sentiment mining.

A. *Variations of sentiment analysis*

Fine-grain sentiment analysis achieves an additional precise level of polarity by breaking it down into any classes, sometimes very positive to very negative. This could be the same as the opinion equivalent of ratings on a 5-star scale. Emotion detection, as the name suggests, figures out a range of emotions instead of positivity and negativity. Examples may include happiness, frustration, shock, anger, and disappointment.

Intent-based sentiment analysis acknowledges movements behind any given text additionally to the opinion of the test. For instance, a web product review expressing frustration concerning dynamical to a battery changing prompt client service to succeed in intent on the resolve that specific issue.

Aspect-based sentiment analysis figures the particular part being positively or negatively mentioned. For instance, a client may leave a review on a product saying the battery life was too short. Then, the system return that the negative sentiment isn't concerning the merchandise as a full, however concerning the battery life.

III. RELATED WORK

Adnan Muhammad Shah et al. in [6] analyzes patient online reviews (PORs) to identify emerging and fading topics and sentiment trends in PRWs during the early stage of the COVID-19 outbreak. An improved latent Dirichlet allocation (LDA)-based topic modeling (topic coherence-based LDA [TCLDA]), manual annotation, and sentiment analysis tool were applied to extract a suitable number of topics, generate corresponding keywords, assign topic names, and determine trends in the extracted topics and specific emotions.

Yingwei Yan et al. in [7] study, a mixed-methods approach involving sentiment analysis and Latent Dirichlet allocation (LDA) topic modeling is designed for mining sheer volume of tweets about Lombok and Bali, generated by nonlocal Twitter users after a series of earthquakes in the two places in August 2018. The findings mainly suggest that people have generally become less negative about Lombok and Bali over time, despite fluctuations in their sentiment polarities' central tendencies. In addition, dissatisfaction about the housing reconstruction progress, tourism recovery status, and living conditions in the affected areas.

Swagato Chatterjee et al. in [8] focuses on the drivers of the helpfulness of textual OHR, for which we have used text-mining techniques to find the sentiment content, polarity, and emotions; we have also used econometric and machine learning techniques to explain and predict its helpfulness. We found that content and title polarity lead to OHRs being less helpful, whereby this negative relationship gets accentuated with higher sentiment content. On the other hand, while negative emotion with low arousal makes OHR helpful, high arousal makes it less helpful. It has also been noted that after controlling for polarity, sentiment, and emotions, longer reviews are less helpful. Higher quantitative rating, recency of OHR, and a reviewer's past expertise make a review more helpful. Additionally, machine-learning techniques have been found to predict 'review' helpfulness marginally better than econometric techniques.

D. Deng et al. in [9] proposed a novel hierarchical supervision topic model to construct a topic-adaptive sentiment lexicon (TaSL) for higher-level classification tasks. It is widely recognized that sentiment lexicon as a useful prior knowledge is crucial in sentiment analysis or opinion mining. However, many existing sentiment lexicons are constructed, ignoring the variability of the sentiment polarities of words in different topics or domains. In TaSL, we solve this issue by jointly considering the topics and sentiments of words. Documents are represented by multiple pairs of topics and sentiments, where each pair is characterized by a multinomial distribution over words. Meanwhile, this generating process is supervised under hierarchical supervision information of documents and words. The main advantage of TaSL is that the sentiment polarity of each word in different topics can be sufficiently captured. This model is beneficial to construct a domain-specific sentiment lexicon and then effectively improve the performance of sentiment classification.

G. Zhai et al. in [10] proposed a sentiment analysis model that can help educators discover the true feelings of students about the course in a timely manner and adjust the teaching plan accurately and timely to improve the quality of education and teaching. Aiming at the inefficiency and heavy workload of college curriculum evaluation methods, a Multi-Attention Fusion Modeling (Multi-AFM) is proposed, which integrates global attention and local attention through gating unit control to generate a reasonable contextual representation and achieve improved classification results.

B. Amina et al. in [11] proposed framework SCANCPECLENS is an initiative to support performance assessment framework for e-government in Pakistan. The research takes into account the opinion of the masses on one of the most crucial and widely discussed development projects, the China Pakistan Economic Corridor (CPEC), considered as a game-changer due to its promise of bringing economic prosperity to the region. The proposed framework suggests using machine learning algorithms to automatically discover the public sentiment from microblogs on the matter nationally as well as internationally. We also present an automated way to create a sentiment lexicon of positive, negative, and neutral words on the subject. To the best of our knowledge, this theme has not been explored for opinion mining before and helps one in effectively assessing public satisfaction over government policies in the CPEC region.

IV. SENTIMENT DETECTION TECHNIQUES

A. Machine Learning Approach

The text classification methods using the ML approach can be roughly divided into supervised and unsupervised learning methods [12, 13]. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is difficult to find these labeled training documents.

B. Supervised learning

The supervised learning methods depend on the existence of labeled training documents. There are many kinds of supervised classifiers in literature [14, 15]. In the next subsections, we present in brief details some of the most frequently used classifiers in SA.

- Probabilistic classifiers
  Naïve Bayes Classifier (NB)
  Bayesian Network (BN)
  Maximum Entropy Classifier (ME)
- Linear classifiers
- Support Vector Machines Classifiers (SVM)
- Neural Network (NN)
- Decision tree classifiers
- Rule-based classifiers

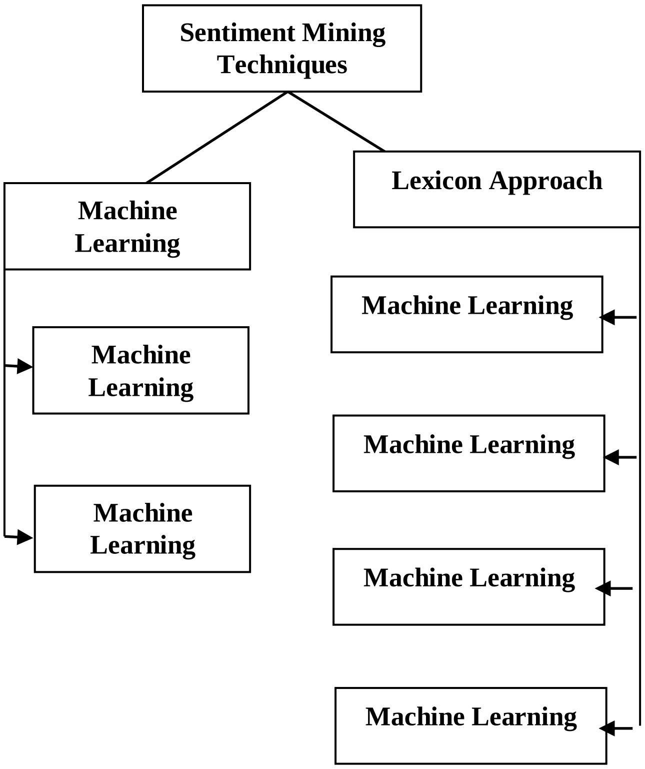

Fig. 1 Techniques of Sentiment mining

C. Unsupervised learning

The main purpose of text classification is to classify documents into a certain number of predefined categories. In order to accomplish that, a large number of labeled training documents are used for supervised learning, as illustrated before. In-text classification, it is sometimes difficult to create these labeled training documents, but it is easy to collect the unlabeled documents [16]. The unsupervised learning methods overcome these difficulties. The concept of weak and semi-supervision is used in many applications. In [17] have proposed a strategy that works by providing weak supervision at the level of features rather than instances. They obtained an initial classifier by incorporating prior information extracted from an existing sentiment lexicon into sentiment classifier model learning. They refer to prior information as labeled features and use them directly to constrain the model's predictions on unlabeled instances using generalized expectation criteria.

The unsupervised approach was used too in [18] to automatically discover the aspects discussed in Chinese social reviews and also the sentiments expressed in different aspects. They used the LDA model to discover multi-aspect global topics of social reviews, and then they extracted the local topic and associated sentiment based on a sliding window context over the review text. They worked on social reviews that were extracted from a blog data set (2000-SINA) and a lexicon (300-SINA Hownet). They showed that their approach obtained good topic partitioning results and helped to improve SA accuracy. It helped too to discover multi-aspect fine-grained topics and associated sentiment.

D. Lexicon-based approach

Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms, which together are called opinion lexicon [19, 20]. There are three main approaches in order to compile or collect the opinion word list. The manual approach is very time-consuming, and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods [21]. The two automated approaches are presented in the following subsections.

- Dictionary-based approach
- Corpus-based approach
  Statistical approach
  Semantic approach

V. FEATURES OF TEXT MINING

Term Frequency: The TF is the count of category-of-words of every category in each document. So the document's term frequency for a category is the occurrence of the words in a single document or article [15].

Document Term Frequency: It is the number of documents in the collection that contain a term. IDF: Inverse Document Frequency is a measure of how much information the word provides, i.e., if it's common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word.

$$IDF(t) = log\left(\frac{N}{n}\right)$$

N represents the total number of documents in the dataset, n represents the number of documents that term t appears

TF-IDF: TF-IDF [16] (Term Frequency-inverse Document Frequency) puts weighting to a term based on its inverse document frequency. It means that if the more documents a term appears, the less important that term will be, and the weighting will be less.

$$TFIDF(t) = TF_t * log\left(\frac{N}{n_t}\right)$$

TF-IDF-CF: As per the Shortcomings of TF-IDF has, introduce a new parameter to represent the in-class characteristics, and authors have called this class frequency, which calculates the term frequency in documents within one class.

$$TFIDFCF(t) = \log(TF_t + 1) * log\left(\frac{N+1}{n_t}\right) * \frac{n_{c,t}}{N_c}$$

The number of documents where term t appears within the same class c document. Nc represents the number of documents within the same class c document.

VI. EVALUATION PARAMETER

In order to evaluate results, there are many parameters such as accuracy, precision, recall, F-score, etc. Obtaining values can be put in the mention parameter formula to get better results.

Actual	**System**	
	True	**False**
Positive	TP	FP
Negative	TN	FN

$$\Pr ecision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$\text{Re}\,call = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$F\_Score = \frac{2 * \Pr ecision * \text{Re}\,call}{\Pr ecision + \text{Re}\,call}$$

In above true positive value is get increases when the model says valid for the sentiment class and really content belong to that class. While if the model says true for a sentiment class and actually content does not belong to the particular class, then true negative counter increases. Additionally, if the model says that content did not belong to a sentiment class and actually content belong to a sentiment class, then the counter of False-positive increases.

VII. CONCLUSION

Web content is a good source to understand a true review or response of any service, product. Many researchers are extracting information from the content in the form of rating, sentiment, feedback, etc. This paper has focused on research work done by scholars who have content available on digital social media sites. The Paper has summarized techniques proposed or adopted by the researcher. Some of the basic text mining features were also detailed with their formulas. Paper has classified sentiment mining work into few classes as well. It was found that feature reduction and transformation play an important role in text sentiment detection work. So in the future, it is expected that scholars can develop a feature filter and machine learning technique for predicting text content sentiment.

REFERENCES

[1] Q. Liu, E. Chen, H. Xiong, C. Ding, and J. Chen., Enhancing collaborative filtering by user interest expansion via personalized ranking, IEEE Transactions on Systems, Man, and CyberneticsPart B (TSMCB), 42(1) 218-233, 201.

[2] X. Qian, H. Feng, G. Zhao, and T. Mei., Personalized Recommendation Combining User Interest and Social Circle, IEEE Trans. Knowledge and Data Engineering, 26(7) (2014) 1487-1502.

[3] L. Spiliotopoulou, D. Damopoulos, Y. Charalabidis, M. Maragoudakis, and S. Gritzalis., Europe in the shadow of the financial crisis: Policy making via stance classification, in Proc. 50th Hawaii Int. Conf. Syst. Sci., (2017) 2835–2844.

[4] S. Hasbullah, D. Maynard, R. Z. W. Chik, F. Mohd, and M. Noor., Automated content analysis: A sentiment analysis on Malaysian government social media, in Proc. 10th Int. Conf. Ubiquitous Inf. Manage. Commun., 30. (2016).

[5] P. C. G. Reddick, A. T. Chatfield, and A. Ojo., A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use., Government Inf. Quart., 34(1) (2017) 110–125.

[6] Adnan Muhammad Shah, Xiangbin Yan, Abdul Qayyum, Rizwan Ali Naqvi, Syed Jamal Shah., Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach., International Journal of Medical Informatics, 149 (2021) 104434.

[7] Yingwei Yan, Jingfu Chen, Zhiyong Wang., Mining public sentiments and perspectives from geotagged social media data for appraising the post-earthquake recovery of tourism destinations., Applied Geography, 123 (2020) 102306.

[8] Swagato Chatterjee., Drivers of the helpfulness of online hotel reviews: A sentiment and emotion mining approach, International Journal of Hospitality Management, 85 (2020) 102356.

[9] D. Deng, L. Jing, J. Yu, S. Sun and M. K. Ng., Sentiment Lexicon Construction With Hierarchical Supervision Topic Model, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(4) (2019) 704-718.

[10] G. Zhai, Y. Yang, H. Wang, and S. Du., Multi-attention fusion modeling for sentiment analysis of big educational data, in Big Data Mining and Analytics, 3(4) (2020) 311-319.

[11] B. Amina and T. Azim., SCANCPECLENS: A Framework for Automatic Lexicon Generation and Sentiment Analysis of Micro-

Blogging Data on China Pakistan Economic Corridor, in IEEE Access, 7 (2019) 133876-133887.

[12] 36. Hanhoon Kang, Seong Joon Yoo, Dongil Han., Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews, Expert Syst Appl, 39 (2012) 6000-6010.

[13] Jonathan Ortigosa-Hernández, Juan Diego Rodríguez, Leandro Alzate, Manuel Lucania, Iñaki Inza, Jose A. Lozano., Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers, Neurocomputing, 92 (2012) 98-115.

[14] 62. Aggarwal Charu C, Zhai Cheng Xiang., Mining Text Data. Springer New York Dordrecht Heidelberg London: © Springer Science+Business Media, LLC,12 (2012).

[15] Cortes C, Vapnik V. Support-vector networks, presented at the Machine Learning, (1995).

[16] Ko Youngjoong, Seo Jungyun. Automatic text categorization by unsupervised learning. In: Proceedings of COLING-00, the 18th international conference on computational linguistics, (2000).

[17] Yulan He, Deyu Zhou., Self-training from labeled features for sentiment analysis, Inf Process Manage, 47 (2011) 606-616.

[18] Fu Xianghua, Liu Guo, Guo Yanyan, Wang Zhiqiang., Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon,Knowl-Based Syst, 37(2013) 186-195.

[19] Hu Minging, Liu Bing., Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04), (2004).

[20] Kim S, Hovy E., Determining the sentiment of opinions. In: Proceedings of the international conference on Computational Linguistics (COLING'04), (2004).

[21] Hatzivassiloglou V, McKeown K., Predicting the semantic orientation of adjectives. In: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL'97), (1997).