Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

A Fish Swarm Based Hybrid Algorithm for Opinion Mining in Healthcare Segment

Vaishali Malik[1], Dr. Nidhi Tyagi[2]

[1]Research Scholar, Shobhit Institute of Engineering & Technology (Deemed to-be University), Meerut (U.P.), vaishalimalik07@gmail.com
[2]Professor, Shobhit Institute of Engineering & Technology (Deemed to-be University), Meerut (U.P.), nidhi.tyagi@shobhituniversity.ac.in

This study uses a unique feature selection approach based on a swarm of tiny fishes hunting for food to identify essential medical terms in blogs and twitter whether the patient approves or disapproves of the therapy. Three most famous classification algorithms were employed to evaluate retrieved features. Fish Swarm Optimization and Tabu Search are combined to solve FSO's local minima issue. To address issues like premature and blind search, we apply alternative computing models in similarity computation based on features and augment the artificial fish swarm method with a Tabu table that has a memory function. Experimental findings show that TDFxIDF with kernel PCA features boost classifier accuracy. CART classifier outperforms others. Features TDF×IDF, Kernel PCA, and hybrid FSO-CART perform better with 86.26% accuracy for medical blog dataset. Similarly Features TDF×IDF, Kernel PCA, and hybrid FSO-KNN with 98.23% camera dataset accuracy. From high-dimensional function optimization trials, the new technique has better global exploration and convergence time and may be employed in various optimization applications.

Keywords: Opinion Mining, Twitter, Digital India.

1. Introduction

Opinion Mining (OM) is a kind of natural language processing used to collect people's opinions on a certain subject, item, or service. Aware of both the subjectivity and objectivity of texts, OM organizes the former according to the viewpoints direction of the latter [1]. The use of internet and smartphones are almost to the point where they can be found everywhere and everyone now can post reviews about the services and goods that they used. In this context, "opinion holders" refers to any individual or organization that has formed a certain point of view. Members of the public with a stake in a subject publish reviews, discussions, and blogs on the products or services for which they have an opinion. Users make informed decisions about hospital to choose to, which doctor has better reviews, as well as the amenities available in the facility. The sheer volume of reviews spread over several websites makes it next to impossible to make sense of them all. Opinion mining provides a synopsis of the available

reviews, as well as their polarity, which may be used to form an overall impression of the product or service in question. Review data is retrieved and then assigned a positive, neutral, or negative Sentiment. Methods like clustering and supervised learning are used to categorize the polarity of opinions. There has been a lot of study put into the process of identifying and labelling emotional states, and the literature reviews various different methods [2]. It was determined how well various feature extraction strategies and classification algorithms performed when applied to the task of categorizing camera review content. All camera reviews are based on information found on the Amazon website. The TDFIDF is used for feature extraction in camera evaluations. PCA and kernel PCA are used to alter the features. It examines the effectiveness of the Nave Bayes, K-Nearest Neighbor, and CART classification algorithms.

In the context of research, "opinion mining" refers to the practice of sifting through vast amounts of user-generated information to find relevant data. Information gleaned from customers' feedback may be formally organized and processed by a computer for further use. While there have been a few of research on the topic of subjective text identification in the last two decades, the real progress in the field didn't start until the explosion of Web 2.0. Different forms of online material gave rise to new challenges and possibilities [3]. Thanks to developments like Web 2.0 and the meteoric emergence of social networks, the daily active user base of the internet has exploded in recent years. Users of social networking sites like Facebook and Twitter may keep in contact with one another by composing and sending private messages, uploading photos, and making other profile updates. Websites like epinion.com, yelp.com, and tripadvisor.com allow users to voice their opinions on various services and products. User-generated content like blogs, comments, reviews, wikis, photographs, etc. is created, creating a big pool of data from which to draw. Due to the fact that it reflects the user's perspective on a given issue, user-generated content may be a valuable resource for web mining. The process of mining for user sentiments on a given subject is called opinion mining, or sentiment analysis. Any current event, product, film, place, hotel, etc., might serve as a topic. Opinion mining entails collecting, organizing, and analyzing opinions from many sources. There has been a recent upsurge in study in this area. As well as analyzing consumer input on products, opinion mining may be used to make informed choices on their behalf. Retrieving opinions refers to the process of accumulating review site opinion content. Content such as reviews, blogs, tweets, micro-blog comments, and online comments all include elements of opinion language. Other resources, such as dictionaries, word lexicons, word senses, etc., may also be used for sentiment analysis [4]. There are specialized areas of machine learning dedicated to doing evaluations of sentiment and mining for people's views. Given the prevalence of user-generated content on the web nowadays, they are more crucial. However, this is still a challenging problem to solve since human speech is not well organized. The difficulty of a computer understanding the meaning of a particular statement. While this may be true, the value of sentiment analysis increases daily. Sub-fields of online content mining include Natural Language Processing and Opinion Mining, which deal with the polarity of client comments on items. The Opinion Mining object model is shown in Figure 1.

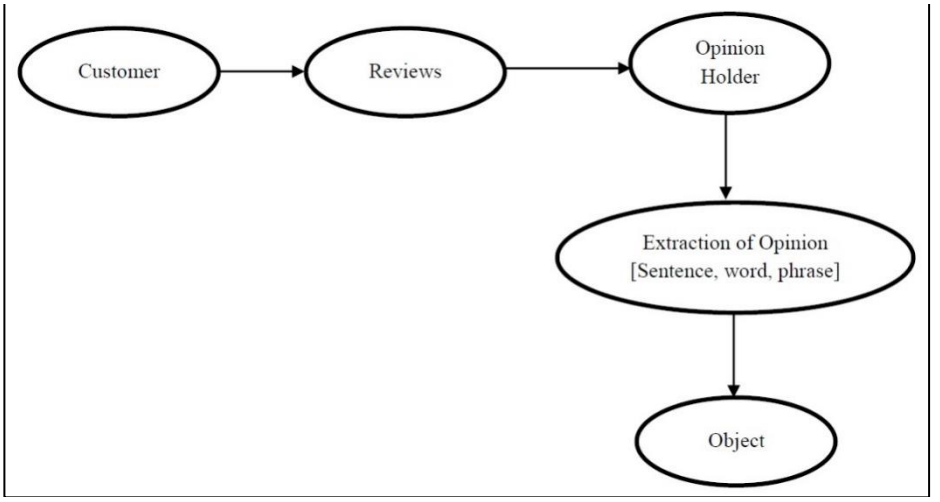

Fig. 1. Opinion Mining object model diagram

The first problem is a subjective phrase that may be interpreted positively under some circumstances and negatively under others. The second is that there is no one universal style in which people expresses their views. Almost all current methods of text processing rely on the assumption that small differences across data sets do not significantly affect meaning. On the other hand, "the picture is excellent" cannot be compared to "the image is not good" in sentiment analysis. It's possible that humans' sentences are internally inconsistent. Whether good or negative, evaluations always include feedbacks, and this must be processed via a sentence-by-sentence examination of the text. The same may be found in plenty in the digital environment in the shape of blogs, reviews, etc. Features selection is a fundamental procedure in opinion mining and sentiment assessments [5]. The process of opinion mining and sentiment assessments is shown in Figure 2. In many ways, the task of opinion mining is the same as that of information retrieval. The aim is not to collect cold, hard facts, but rather to get a sense of the text's polarity and how people feel about it. Due to the subjective and context-sensitive character of views value, identifying polarity phrases that represent their authors' sentiments about a certain issue remains a challenging challenge [6].

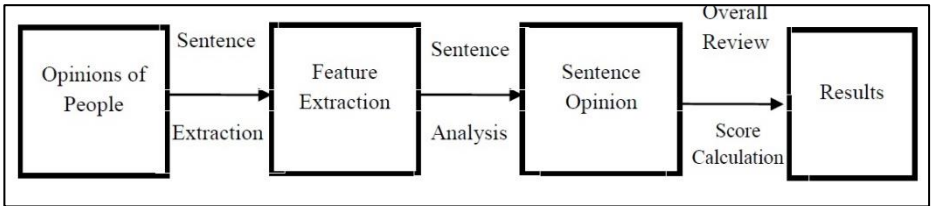

Fig. 2. Analysis of Opinions and Emotions as a Process

Natural language processing (NLP) of the sentiment analysis kind monitors the public's feelings about a product or issue. One of many possible applications for sentiment analysis. For instance, market research aids to the success of marketing campaigns or the introduction of new items by revealing which edition of products or services is most loved and even detecting the demographic that favours certain elements. With Sentiment Analyses, there are

several challenges to overcome. Typically, review datasets consist of public opinions expressed in natural languages with associated numerical scores describing the evaluation, and are used to evaluate sentiment analysis algorithms due to their usefulness in the sectors of marketing and reputation management. Because of this, a corpus of 2000 annotated English movie reviews for sentiment is often used to test different approaches to sentiment analysis. Sentiment analysis methods are shown in Figure 3.

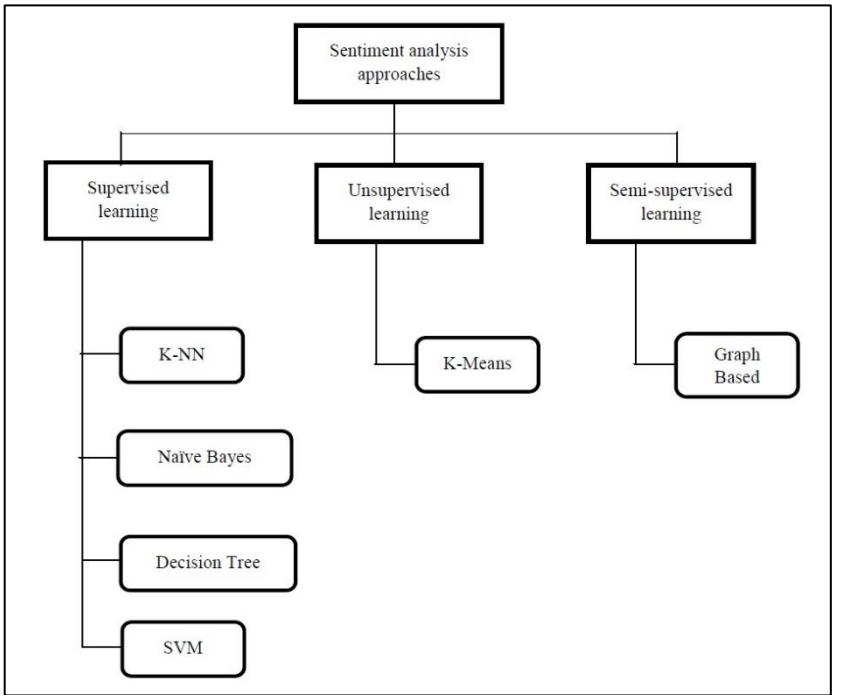

Fig. 3. Methods followed in Opinion Analysis

2. Literature Review

In paper [7] authors investigated aspect-level OM and proposed an unique syntactic based approach that combines syntactic dependence, aggregated score of opinion terms, SentiWordNet, and aspect table for OM. Research was done on how to best use restaurant ratings. Manually collected data on restaurant ratings and comments was labelled after being compiled online. On the annotated test dataset, the proposed method achieved an overall accuracy of 78.04%. When compared to a method that relied on a Part-Of-Speech tagger to extract features, the proposed method was shown to be 6 percent more accurate on the annotated test dataset.

A method based on a naive Bayes classifier was suggested in paper [8] to identify user attitudes in movie reviews. The field of opinion mining is dissected, along with the methods of sentiment analysis and the uses to which they might be put. We have built the proposed method, evaluated its performance, and made recommendations for further improvement.

In paper [9] authors centred on a classification of opinion mining methods that may be utilised to express a user's opinion (positive or negative) at different granularities. Accurate opinion forecasting techniques now allow for the extraction of feelings from the web and the prediction of the preferences of online customers, which might be useful in the field of marketing research. Thoughts are so valuable that if you need to make a choice and want to hear the opinions of others, much study has been done on the processing of opinions or feelings. The user's perspective was crucial, but the business also benefited from it.

Taking a domain-oriented approach, [10] created a collection of domain-specific resources that record important information about how individuals express themselves in a certain domain. These tools are inferred mechanically from a corpus of annotated texts. Experiments comparing the method to other state-of-the-art, domain-independent algorithms were conducted on data from three distinct domains (user-generated evaluations of headphones, hotels, and autos). The findings reaffirmed the significance of the domain in the development of reliable opinion extraction systems.

Authors of [11] added the capability to handle large-scale restrictions to Latent Dirichlet Allocation (LDA), a common topic modelling tool. Then, two cutting-edge strategies for automatically extracting both kinds of constraints were developed. As a last step, we use the constrained-LDA and the extracted constraints to classify product characteristics. In experiments, constrained-LDA exhibited much better results than both the baseline LDA and the state-of-the-art multilevel Latent Semantic Association (mLSA).

To solve the issue of automated sentiment categorization of reviews, authors of [12] used the SentiWordNet lexical resource with many different methods and published their findings. This method outperformed prior implementations of SentiWordNet for review sentiment categorization.

In [13], a new approach to sentiment analysis using clustering was presented. Using a Word Frequency-Inverse Document Frequency (TF-IDF) weighting strategy, voting mechanism, and importing term scores may provide a suitable and consistent clustering result. In comparison to the existing two main techniques, symbolic methods and supervised learning, this one is much superior. The method was efficient, did not need human intervention, and improved performance in sentiment analysis.

In order to effectively extract embedded emotions and avoid being led astray by simply linguistic signals, authors of [14] suggested that Natural Language Processing (NLP) systems, should better model how individuals would assess different states of the environment in situations of interest. Sentences like "Critics say" and "Despite this" contain a wealth of semantic information that can be unlocked using "full semantics" language processing that can describe the interplay between lexical and syntactic semantics and then weave them together with domain knowledge. We have integrated ideas, knowledge, language processing, and opinion mining using domain models based on the Radical Construction Grammar-based COGPARSE parser to get these insights.

3. Proposed Approach

3.1 Methodology

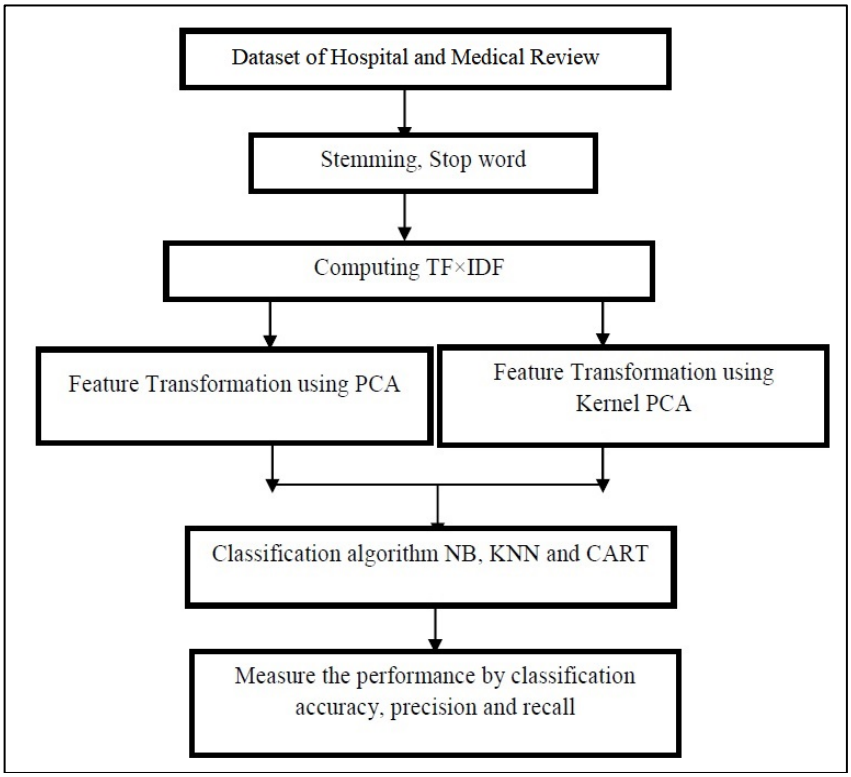

Fig. 4. The procedure used is outlined here in a flowchart

3.2 Feature Extraction

Stemming: The term "stemming" refers to the etymological beginnings of words. For instance, the words talk, speaking, and speaks all come from the same root, which is speak. Morphological variations of words are often treated as equivalent for IR purposes since they have equivalent meanings. Therefore, several algorithms (called stemmers) were developed to reduce a word to its root form. So, their stems rather than their full forms represent the most important words in a query or document. This means that the many forms of a word may be consolidated into a single, more accurate representation, and that the number of unique terms needed to characterise a given body of text can be cut down significantly.

Stop Word: There are two primary motivations for compiling a list of terms that serve no useful function in retrieval but are still widely used in written communication. For starters, a well-indexed query and document match may have been founded on similar concepts. That is why it is not a good idea to use a search query that includes the terms "be," "the," and "your" to find relevant documents. As noise, these insignificant words harm document retrieval efficiency by making it harder to tell which documents are important and which aren't. Additionally, it should lower the size of inverted files by 30–50%.

3.3 Feature Transformation

As a document-specific measure of term relevance, Term Frequency (TF) represents the number of times each word appears in the text. A corpus refers to a group of related texts that are being studied. The literature suggests a variety of word weighting methods. The word weights in a text are the components of a vector space model. One way to measure a word's significance inside a corpus or collection is using its TF-IDF weight (Term Frequency-Inverse Document Frequency). The weight of TF-IDF grows as a word occurs more often in the text, whereas it decreases as a word appears more frequently in the corpus. Search engines rely heavily on TF-IDF weighting and its variants to determine how relevant a page is to a user's query. Inverted document frequency (IDF) is calculated by determining how many documents in the collection being searched include the phrase of interest. A term's IDF is considered to be zero if it appears in every file in the set.

$$\mathrm{idf_i} = \log\frac{|D|}{|\{j:t_i \epsilon d_j\}|} \quad (\text{SEQ "equation" \n * MERGEFORMAT 1})$$

With |D|: cardinality of D, or the total number of documents in the corpus |j: t_i d_j|: number of documents in which the word t_i occurs (document frequency) ($n_{i,j} > 0$). The TF count is adjusted to avoid a bias toward lengthier documents in order to determine the significance of the word t_i inside the document d_j.

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\Sigma_k \; _{,j}} \quad (2)$$

here $n_{i,j}$ is the number of times that the considered word (t_i) appears in document d_j, and the denominator is the total number of times that any term appears in document d_j, often known as the size of the document $|d_j|$.

3.4 Feature Selection

Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA) is the method that is used the most often for feature selection. PCA is a linear dimensionality reduction approach that projects higher dimensional information along major axes that are picked by computing variance. It is a method that is usually employed. Several correlative approaches that exhibit better capability than principal component analysis in various application domains have been proposed as a result of improvements made to principal component analysis.

4 Results and Discussion

The accuracy of the Naive Bayes, KNN, and CART classification methods for reviews is compared and contrasted. Feature extraction utilising simply TDF-IDF, feature extraction using TDF-IDF and PCA, and feature extraction using TDF-IDF and kernel PCA are all subjected to experimentation. The accuracy of OM is evaluated using statistical measures like as recall and precision, as well as the F1 measure. One method of judging something is called "accuracy," and it may be defined as the proportion of people whose feelings are appropriately

anticipated. Recall and accuracy are the two typical performance measures that are compared in order to determine how successfully the quality of the results has been measured. One way to think about recall is as the fraction of really upbeat feelings that can be reliably recognised.

Table 1. Accuracy in Classification, Precision, and Recall Utilizing Medical Blogs and data medical reviews from social media

	Classification Accuracy (%)	Precision (%)	Recall (%)
FE using TDFxIDF – NB	74.56	74.569	74.125
FE using TDFxIDF and PCA- K-NN	74.66	74.66	74.667
FE using TDFxIDF and PCA- CART	78.45	78.12	78.22
FE using TDFxIDF – K-NN	73.70	73.76	73.79
FE using TDFxIDF and PCA- NB	75.51	75.557	75.557
FE using TDFxIDF and kernel PCA - K-NN	74.66	74.52	74.527
FE using TDFxIDF and kernel PCA- CART	79.12	79.196	79.12
FE using TDFxIDF – CART	77.32	77.38	77.122
FE using TDFxIDF and kernel PCA- NB	76.22	76.265	76.211

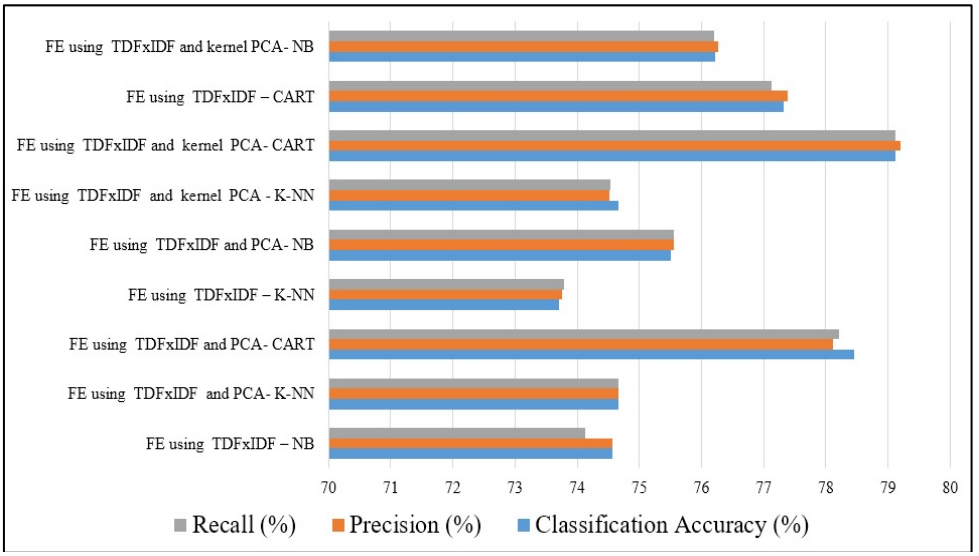

Fig. 5. Classification Accuracy, Recall and Precision for Medical Reviews

The classification accuracy, precision, and recall are each shown in table 1 and figure 5 for the Medical Blogs dataset. Based on the findings of the visual analysis, it is clear that CART achieves a higher level of classification accuracy than NB and K-NN. In a similar vein, the performance of FE models built with TDFxIDF and kernel PCA-CART is superior than that of FE models built with TDFxIDF-NB and FE models built with TDFxIDF-K-NN. According to the findings, the FE that makes use of TDFxIDF and kernel PCA-CART performs more accurately than the FE that makes use of TDFxIDF-CART and by 0.85% more accurately than the FE that makes use of TDFxIDF and PCA-CART.

5 Conclusion

Discovering the thinking of persons is a significant part of the function that information-gathering activities play. Because we are unable to use information technology to understand the perspectives held by others, the availability and rising popularity of resources that are abundant in opinions, such as online reviews and blog sites, has led to an increase in the number of changes and obstacles that have arisen as a result. For the purpose of categorising medical reviews, several feature extraction strategies and classification algorithms were analysed for their effectiveness. The medical reviews were gathered from any social page that is related to the medical segment. The TDF-IDF algorithm is used to extract features from the reviews. PCA and kernel PCA are used to do the transformation on the features. The characteristics may be categorised as positive or negative based on the results of the CART, Naive Bayes, and K-NN classifiers. The experimental findings indicate that the classification accuracy of the classifiers may be improved by extracting features using TDF-IDF with kernel principal component analysis (PCA). Based on the findings, it has been determined that the CART method has a greater classification accuracy than other classifiers.

References

1. Kumar, Akshi, Prakhar Dogra, and Vikrant Dabas. "Emotion analysis of Twitter using opinion mining." In 2015 Eighth International Conference on Contemporary Computing (IC3), pp. 285-290. IEEE, 2015.
2. Anjaria, Malhar, and Ram Mohana Reddy Guddeti. "Influence factor based opinion mining of Twitter data using supervised learning." In 2014 sixth international conference on communication systems and networks (COMSNETS), pp. 1-8. IEEE, 2014.
3. Baly, Ramy, Gilbert Badaro, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Wassim El-Hajj, Nizar Habash, and Khaled Shaban. "A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models." In Proceedings of the third Arabic natural language processing workshop, pp. 110-118. 2017.
4. Tavoschi, Lara, Filippo Quattrone, Eleonora D'Andrea, Pietro Ducange, Marco Vabanesi, Francesco Marcelloni, and Pier Luigi Lopalco. "Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy." Human vaccines & immunotherapeutics 16, no. 5 (2020): 1062-1069.
5. Bouazizi, Mondher, and Tomoaki Ohtsuki. "Opinion mining in twitter how to make use of sarcasm to enhance sentiment analysis." In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 1594-1597. 2015.
6. Barnaghi, Peiman, Parsa Ghaffari, and John G. Breslin. "Opinion mining and sentiment polarity on twitter and correlation between events and sentiment." In 2016 IEEE second international conference on big data computing service and applications (BigDataService), pp. 52-57. IEEE, 2016.
7. Sharma, Bhavna, Ravi Prakash, Shailesh Tiwari, and K. K. Mishra. "A variant of environmental adaptation method with real parameter encoding and its application in economic load dispatch problem." Applied Intelligence 47, no. 2 (2017): 409-429.
8. Shukla, Richa, Bramah Hazela, Shashwat Shukla, Ravi Prakash, and Krishn K. Mishra. "Variant of differential evolution algorithm." In Advances in computer and computational sciences, pp. 601-608. Springer, Singapore, 2017.
9. Prakash, Ravi, Suresh Kumar, Chandan Kumar, and K. K. Mishra. "Musical password based biometric authentication." In 2016 International Conference on Computing, Communication and

Automation (ICCCA), pp. 1016-1019. IEEE, 2016.
10. Meduru, Manogna, Antara Mahimkar, Krishna Subramanian, Puja Y. Padiya, and Prathmesh N. Gunjgur. "Opinion mining using twitter feeds for political analysis." Int. J. Comput.(IJC) 25, no. 1 (2017): 116-123.
11. Zervoudakis, Stefanos, Emmanouil Marakakis, Haridimos Kondylakis, and Stefanos Goumas. "OpinionMine: A Bayesian-based framework for opinion mining using Twitter Data." Machine Learning with Applications 3 (2021): 100018.
12. Gokulakrishnan, Balakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath, and AShehan Perera. "Opinion mining and sentiment analysis on a twitter data stream." In International conference on advances in ICT for emerging regions (ICTer2012), pp. 182-188. IEEE, 2012.
13. Khan, Farhan Hassan, Saba Bashir, and Usman Qamar. "TOM: Twitter opinion mining framework using hybrid classification scheme." Decision support systems 57 (2014): 245-257.
14. Zhou, Xujuan, Enrico Coiera, Guy Tsafnat, Diana Arachi, Mei-Sing Ong, and Adam G. Dunn. "Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter." (2015).
15. Bing, Li, and Keith CC Chan. "A fuzzy logic approach for opinion mining on large scale twitter data." In 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, pp. 652-657. IEEE, 2014.
16. Beck, Tilman, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. "Investigating label suggestions for opinion mining in German Covid-19 social media." arXiv preprint arXiv:2105.12980 (2021).