Opinion Mining On Web-Based Communities Using Optimised Clustering Algorithms

Suhaib Bin Younis[1], Dr. R. Naresh[2,*]

[1]B.Tech-student, CSE, SRM Institute of Science and Technology.

[2*]Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603 203.
E-Mail: sm4695@srmist.edu.in, nareshr@srmist.edu.in
*Corresponding Author: R. Naresh

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract- People like to consume content by downloading it from the internet, together with contribution and creation of new content, thanks to the growth of social networks and the Web. People also turn out to be more anxious to communicate and impart their insights on the web with respect to everyday exercises or worldwide issues. Sentiment analysis and opinion mining are rapidly evolving as a result of the rapid growth of web-based platforms such as Facebook, Reddit, Twitter, and others. This is from the world of data mining and NLP, especially in the mining of web data and the mining of texts. Why has sentiment analysis been more popular in recent years, which is also known as opinion mining? Whenever we attempt to choose to buy an item, we are probably going to hear the thoughts of companions or family members and do a few overviews before we buy the item. Opinions are now undoubtedly the main drivers of our behavior. The positive, negative, or neutral polarities are consistently seen in the opinions. Opinion mining can be used to discover opinion polarity classification, subjectivity identification, spam detection, and opinion summarization using emotional classification.

Index Terms- Clustering; Opinion Mining; Social Media; Machine Learning;

I. INTRODUCTION

In this period of the computerized world, mining individual's assessment is pivotal to uncover the helpful bits of knowledge and their conclusions with respect to a particular substance. It is a typical practice that with regards to decision making, people like to search out other assessments. The growth of new media information makes sentimental examination an important area to explore other people's judgments. Sentimental Analysis (also called Opinion Mining) is a research discipline that examines people's beliefs, desires, assessments, perceptions, and emotions, as well as their attributes and aspects.

For what reason do we need opinion mining? From the business point of view, proposal and assessment on an item/product can be controlled by opinion investigation to the trader and the client. From the political point of view, political individuals need political data to win the political elections, and knowing the polarity of vote banks is crucial. From a public welfare point of view, sociopolitical opportunities that raise concerns and the need to revise the public safety definition.

The study of sentimental analysis, also known as opinion mining, is used consistently mutually. The sentimental research, though, is about determining branches with regard to whether the stuff is the good or the bad opinion. While the study of subjectivity relates more to opinion mining; than whether a text or document provides opinions.

II. LITERATURE REVIEW

Examination of research sentiments plans to generate opinions from data collected in documents and websites[1-6]. This analysis is utilized to survey client reviews of a product, proposal framework, or be it matters of identifying with business activities. The reason for arranging estimations or suppositions is to distinguish and do the polarity detection of different sorts of conclusions.

In comparison sentiment analysis is divided into two classes:

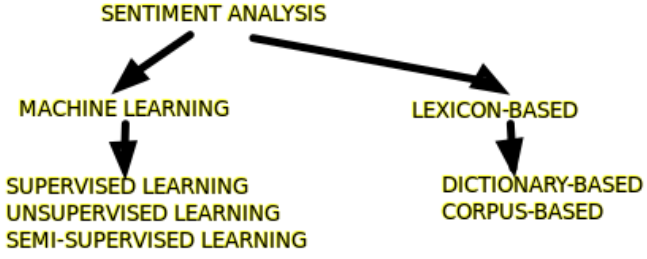


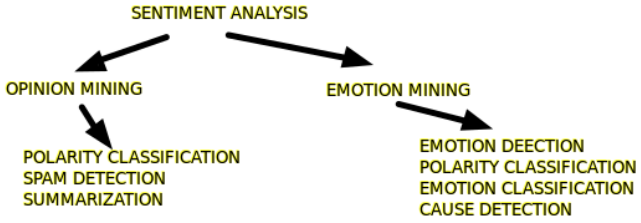
Fig .1. Sentiment Analysis

1) The semantic orientation of terms and phrases is expected to decide the polarity of a record, according to Lexicon Analysis. Implementations, on the other hand, refer to lexicon research and represent little thought regarding the context under consideration.

2) In order to explore text orientation, machine learning includes the construction of models from labeled testing datasets. Studies on the sort of subject are conducted for this type of system.

The two ways of performing analytical sentiments approaches are lexicon-based and machine learning techniques.

There are three key layers of goal review in opinion mining: document, sentence, and aspect level. Aspect-based sentiment analysis is more systematic than sentence-based and document-based sentiment analysis across the three layers of sentiment analysis.



According to the authors of [7-11], sentiment analysis can be applied to the phrase level rather than the document level to calculate the polarity of sentiments for each term and the results of these techniques can be a better sentiment towards the product. According to the authors of [12-18], they got satisfactory results by achieving 91% accuracy in aspect extraction and 93% accuracy in sentiment detection by automatically extracting aspects from the available text and their corresponding orientation with no predefined aspect categories[19-21]. Their model used two LSTMs for the task of aspect extraction and polarity detection respectively.

III. PROPOSED SYSTEM

It's difficult to say which classification strategies would produce the best results based on the research done so far. Various methods, based on a variety of methodologies and algorithms, have been tested and tried on a variety of datasets, yielding varying results. As a result, a general framework for text sentiment analysis was suggested, which uses an unsupervised machine learning approach, as shown in the figure 2.

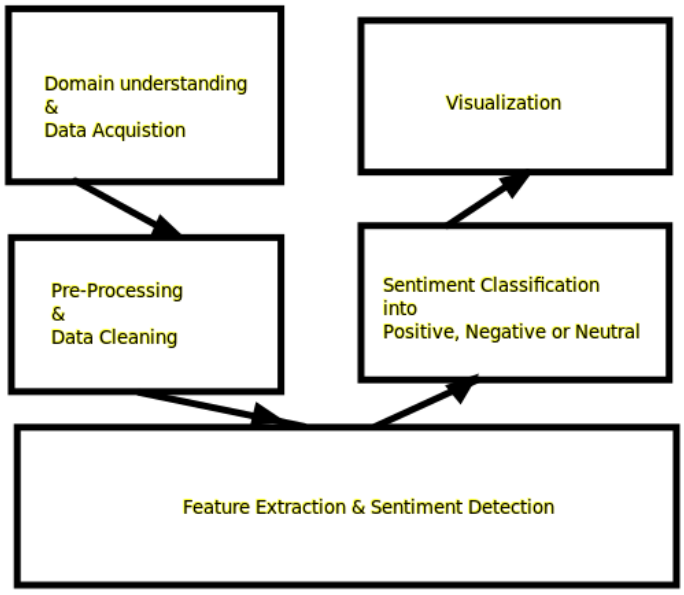


Fig.2 system Architecture

A. Data Collection

The text data is collected from various sources like HTML, PDFs, DOCX, etc

B. Pre-Processing
At this point, text data is pre-processed with linguistic tools such as stopword elimination, speech tag component (POS), tokenization, lemmatization, etc.

1. Tokenization
Tokenization is the process of breaking down a text stream into tokens, which are words, terms, or other significant components. The aim of tokenization is to look at the meaning of individual words in a sentence. Text data is nothing more than a text translation or a string of characters at the start. During the retrieval process of data, words from the data set are required. As a consequence, a parser to manage record tokenization will be needed. In this case, there are a few issues to be addressed, such as punctuation cancellation and assorted characters including parentheses, hyphens, and so on. The primary motivation for tokenization is to find relevant words. A number of issues may occur if abbreviations and acronyms are not translated to standard systems.

2. Discarding Stop Words
Prepositions, articles, and pronouns are examples of well-recognized stop words that are unlikely to help with text mining. These terms are not needed for text mining applications since each text archive is associated with them. These terms were omitted from the equation. We may use any collection of terms for this purpose. It also eliminates text data and speeds up process execution. For instance, "are", "I", "you".

3. Transforming the cases
After that, all words that have been obtained are converted to lowercase letters.

4. Stemming
The aim of this procedure is to extract root words from the dataset's words. The idea is to obtain a root word to avoid counting errors when extracting syntactic features. If we don't steam, the distance score from the grammar highlight won't be 0. For example, if a pair of sentences have the same root word for both of their words but only differ from a postfix, infix, or prefix, the distance score from the grammar highlight won't be 0.

5. Scoring Words
TF-IDF or TFIDF is statistical metric used to represent the importance of a term in a report or corpus when extracting material. In text mining, this is commonly used as a weighting element. The TF-IDF values rise in proportion to the number of times a term appears in an archive. TF-IDF is one of today's most well-known time-weighting schemes, with TF-IDF responsible for 83 percent of the text-based proposal layout in the machine library. The equations for these quantities are below.

Term Frequency, tf:
tf(t,d) = count of word t in the set of words d / total number of words in d

Document Frequency, df:
df(t) = occurrence of t in documents

Inverse Document Frequency, idf:
idf(t) = N/df

in case of a large corpus, the IDF value crashes, to avoid the effect we take the log of idf
idf(t) = log(N/(df + 1))

now a variation of tf-idf can be set which will be the right measure to evaluate how important a word is any text or document:
tf-idf(t,d) = log(N/(df + 1))* tf(t,d)

where,
t -> term / word
d -> document / set of words
N -> count of corpus

C. Feature Extraction & Sentiment Detection
To build any robust classifier, feature extraction is a must when using machine learning approaches. The quality of the model depends on the quality of features. During sentiment detection, reviews are examined for

subjectivity. Sentences that contain subjective expressions are kept and those sentences that contain objective facts are rejected.

D. Sentiment Classification
It is the heart of the system. For any text data that is pre-processed and on which sentiment detection has been done, two approaches can be used when using machine learning i.e supervised learning and unsupervised learning to classify the data.

E. Visualization
When the reviews are classified, the system output is presented to the user with the help of visualization tools that visualize data in a variety of formats, including bar charts, line diagrams, pie charts, and other visual representations.

IV. EXPLANATION OF THE PROPOSED MODEL

To begin with the classification of sentiments, we can further partition classification into semi-supervised, supervised, and unsupervised learning strategies. Any algorithm that learns from the training dataset to accomplish the ideal yield is known as a supervised learning process, while then again if the algorithm is left all alone to discover and show the structure in the information that is interesting, then it is called an unsupervised learning process. Semi-supervised learning processes are based on an algorithm that integrates both supervised and unsupervised.

For the definition of opinion, we represent it as a quintuple as in formula:

$$(E_a, A_{ab}, O_{abcm}, H_c m_m)$$

Where
E_a – any entity ‘a’
A_{ab} – ‘b’s aspect towards entity ‘a’
H_c – any opinion holder ‘c’
m_m – moment ‘m’ when opinions are shared
O_{abcm} – ‘a’s opinion on ‘b’s aspect from holder ‘c’ at moment ‘m’

Clustering and Evaluation

In this paper, we have proposed a framework that approaches the issue with an optimized clustering algorithm. Clustering means assigning a list of items to a cluster because items are more identical to each other in a similar cluster than to items in a different cluster. It is a key concept for the exploration of data and a typical tool for data mining research. The correct algorithm and parameter settings for the clustering depend on the individual data sets and the intended use of results. The data preprocessing and model parameters need to be adjusted before the result is expected to have some ideal properties.

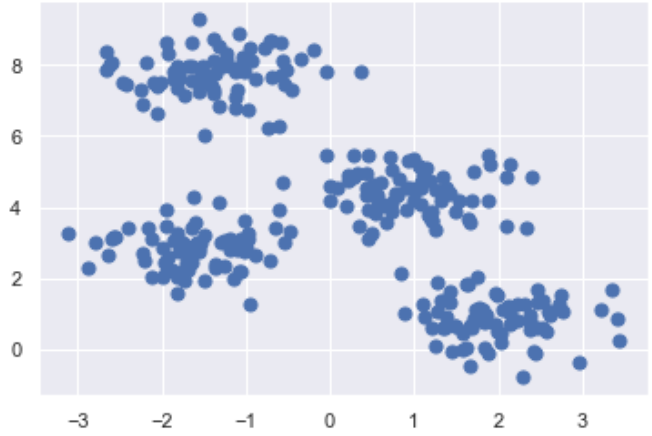
Types of clustering:

a. Hard Clustering: Either every data point has a location with a group or not in hard clustering.

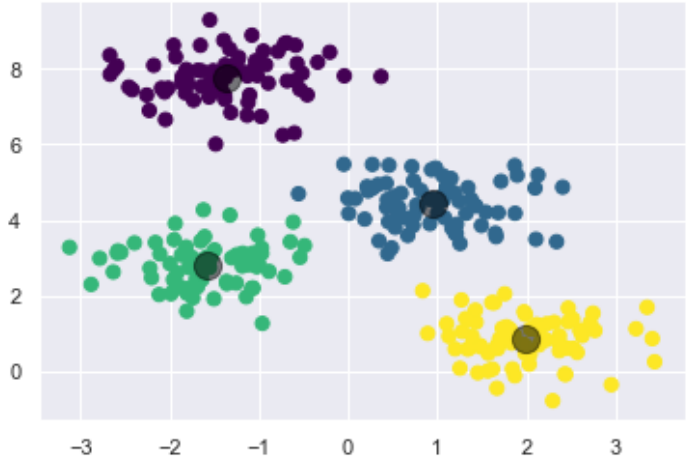
b. Soft Clustering: In soft clustering, a probability or chance of this data highlight in such categories is assigned instead of each data point in a separate cluster.

Clustering is a subjective topic which means that there are plenty of ways to do it. There are currently over 100 algorithms known for clustering. In our approach, we are using the centroid models.

The below images shows data before clustering.



After Clustering the data looks like this. Different colors represent different clusters.



Clustering algorithms in which the definition of similarity depends on the proximity of a data point in the cluster core and are iterative in nature, are known as centroid models. One of the algorithms in this category is the k-means clustering algorithm. The cluster number needed for this purpose must already be referred to in these models, making it essential to be aware of the dataset beforehand. These models work to locate the local optima iteratively.

The algorithm works in the following way:

1. First we take k points(means), which can be random.

2. We arrange every point to its nearest mean and we update the mean's directions, which are the midpoints/averages of the things ordered in that means up until now.

After each iteration, we get k centroids with each point being assigned to its cluster. The cost function is defined as- a summation of Euclidean distance of each point in its cluster center and this is summed over k clusters.

$$\text{Cost} = \sum_{i=1}^{K} \sum \text{Dist}(C_i, x)^2$$

3. We repeat this cycle for a given number of iterations and toward the end, we have our clusters.

KNN is a distribution-free statistic and lazy algorithm where there are no propositions on the basis of the distribution of the database and there is no clear level of preparation prior to classification. The object is organized by the votes of its neighbors and allocated to the most commonly used class of its k neighbors.

Again it should be noted that only machine-learning methodology is consistent with the current general architecture.

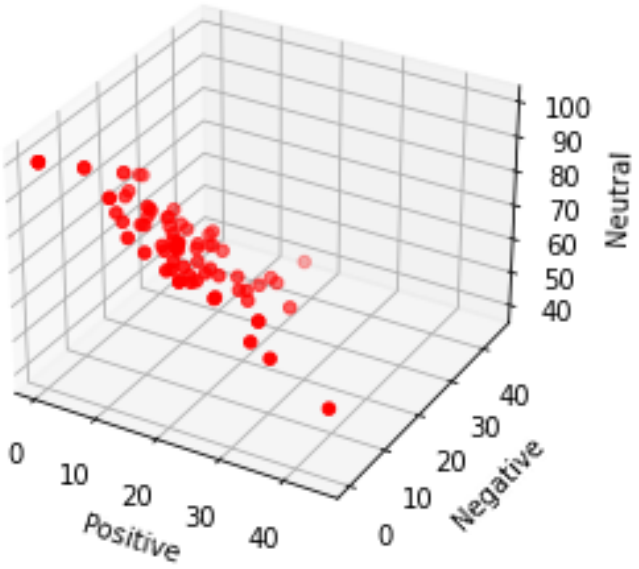
V. RESULTS OBTAINED

This algorithm was tested on reviews of the recent movies on IMDB. The links of the web pages are used to grab the main review and rating from the website. For the testing purpose, 25 reviews each from four movies (Tom and Jerry, Coming 2 America, Raya and the last dragon and Army of the dead) were collected and the following results were found when creating 4 clusters. The system info was as given below:

System Info:

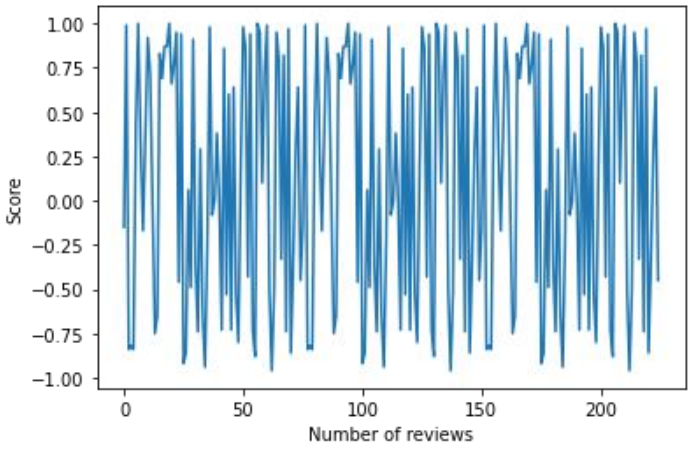
	0
0	Linux-5.4.0-67-generic-x86_64-with-debian-bullseye-sid
1	Python 3.7.6 (default, Jan 8 2020, 19:59:22) \n[GCC 7.3.0]
2	NLTK3.5
3	bs44.9.3
4	Pandas0.25.3
5	sklearn0.24.1
6	np1.20.1
7	Pandas0.25.3
8	sklearn0.24.1
9	re2.2.1

The sentiments obtained can be observed from the following 3D plot.

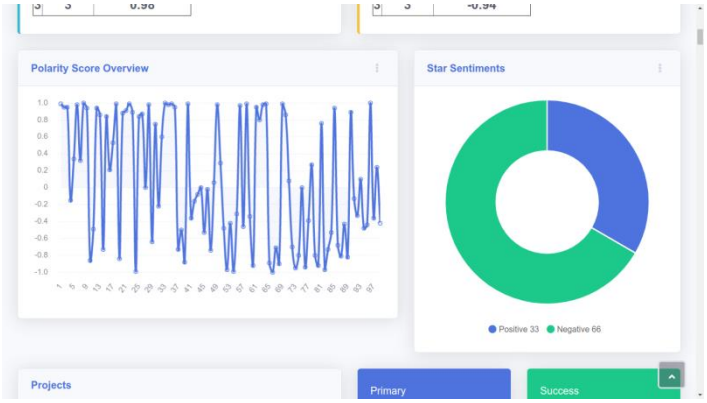


.From the clusters formed we had the following observations:

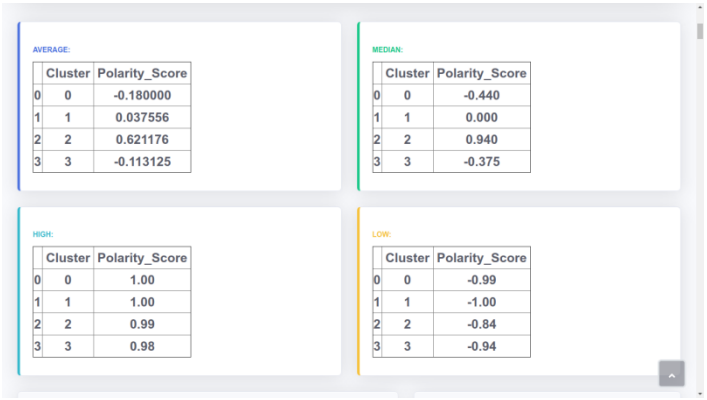
1. On average all clusters had at least one positive sentiment and a positive score of more than 0.002143.
2. All clusters had at least a negative sentiment.
3. The highest score was 1.00 and the lowest was -0.99.



Before prediction was made the total numbers of positive sentiments according to star rating were just 33% while the rest were negative. After the model predictions were made, the positive sentiments rose to 45%.



From the four clusters obtained each cluster was checked for its average, median, high and low values.



AVERAGE:

	Cluster	Polarity_Score
0	0	-0.180000
1	1	0.037556
2	2	0.621176
3	3	-0.113125

MEDIAN:

	Cluster	Polarity_Score
0	0	-0.440
1	1	0.000
2	2	0.940
3	3	-0.375

HIGH:

	Cluster	Polarity_Score
0	0	1.00
1	1	1.00
2	2	0.99
3	3	0.98

LOW:

	Cluster	Polarity_Score
0	0	-0.99
1	1	-1.00
2	2	-0.84
3	3	-0.94

All reviews from the training dataset were given a positive or negative sentiment based on their star ratings and then their closeness was studied to the test dataset. If a star rating was 5 or more than 5 it was considered positive or else negative. The dataset was split into 30% training data and 70% testing data. For each review its closeness to positive, negative and neutral was calculated by the model and a score was associated to it.
Each review was then assigned a cluster for that text can be classified in clusters.

	Movie	Review	User_Rating	Star_Sentiment	Cluster	Polarity_Score
0	Tom_and_Jerry	Always was a fan of the cat and mouse duo "Tom and Jerry" so for sure I watched the latest big screen movie, and must say it was really good and entertained as the way the animation mixed with live action and people characters was super great. Set in New York City the big apple the tale involves a wedding(only things get elephant like crazy!) And both the cat and mouse take up shop in the luxury loft elegant hotel and their chases entertain! Wrapped around it all is a young lady and hotel lobby worker Kayla(Chloe Grace Moretz) as all become friends and learn about the ins and outs of life! The music score and scenes are well done as it even has avid fans thinking back to memories of the old classic Hanna Barbera days. In the end things stay the same with the classic chase rivals only it's live happily ever after for others. Good tribute and entertaining movie of the classic cat and mouse duo.	9	Positive	2	0.99
1	Tom_and_Jerry	This movie is great, a lot of fun. Enjoyed it very much. Animation is amazing. Cast compliments Tom and Jerry.	9	Positive	2	0.95
2	Tom_and_Jerry	I find more ridiculous and sometimes just pathetic than listening to grown adults breaking down why a kids movie is so bad. News flash - this ISNT made for you. I get so much joy watching my 7yr old daughter laugh and laugh and laugh that i could careless what my own personal opinion is, cause in that reality, the movie	8	Positive	1	0.95

After the model was executed, for each review we have the assigned sentiment and the predicted sentiment in the output along with their closeness in being a positive, negative and neutral sentiment.

	Review	Labeled Sentiment	Predicted Sentiment	Positive	Negative	Neutral	Predicted Score
0	Way too much time passed where this could have made any sense. The original movie was brilliant and they wrecked it for me. quit while you ahead ! Flat out too many cheap CGI effects. Story line sucked ! Watch it if you want to waste your time.	Negative	positive	15.0%	11.0%	75.0%	0.29
1	Boring and very poorly written. They pulled this story out of a hat and soiled Akeem's character, as well as Semmi...Really? Semmi got Akeem so messed up that he didn't realize he was basically raped by some trashy broad from the club? This tarnishes the love story of the original film.The love story in this one is forgettable. I also find it interesting that this was another forced attempt by Hollywood for female empowerment. Of course the girl was gonna be allowed to be queen when it's all said and done. Do I have a problem with that? No. But the forced narrative is an issue to me.I'll say this though, to me the most entertaining scenes all involved Wesley Snipes. I found his character to be hilarious. I couldn't help but laugh everytime he walked in. Might just be me though! 😂	Negative	negative	9.0%	13.0%	78.0%	-0.73
2	TOM & JERRY from the genius legend Hanna Barbara made a legend Cartoon movie and this film is absolutely funny from Warner Brothers. You will enjoy splendid direction and great action, comedy and entertainment from the beginning to end. A must see film to forget your problems.	Positive	positive	33.0%	8.0%	60.0%	0.94
3	I don't know how the movie get 8.4....i was getting crazy couldn't wait till the movie is over, that much it's was boring and I didn't understand at all 3 hour what is it all about, from first til lthe end.	Negative	negative	0.0%	12.0%	89.0%	0.57

Finally in the end, we have the feature words from each clusters which complete our feature extraction.

FEATURE WORDS FROM 4 CLUSTERS

	0	1	2	3	4
0	story	film	marvel	boring	movie
1	movie	like	disney	good	thing
2	tom	jerry	cartoon	movie	great
3	movie	original	funny	terrible	jones

VI. APPLICATIONS OF OPINION MINING

Opinion Mining is a valuable instrument in which public opinion or mood towards some company or party is of importance for its effectiveness. Mining opinion takes into account the content of web pages, tweets, letters, emails, and other streams of information to recognize continuing ideas that reveal consumer's optimistic or unfavorable feelings. It helps you to find out what your customers want and what they do not like about your product or service. Such information can be used to solve problems, boost customer support and organize new efforts for the growth of any business. This allows you to develop goods and services that fit your needs and the needs of your client. If you have the right software you can automatically conduct opinion mining on nearly any kind of data that requires very little to no human involvement.

VII. FURTHER DISCUSSIONS AND OBSERVATIONS

While technology is becoming more and more advanced, AI-driven models still remain imperfect. Part of the problem is that there exists a ton of challenges faced when performing opinion mining, such as dealing with spam, detecting fake reviews, and decoding sarcasm. People speak in diverse ways. They can speak in slang or they can use an emoji to express their feelings. The internet and social media that we know today are actually culturally and socially very complex and hard to decode. In our approach, we tried to solve this problem with unsupervised machine learning techniques using clustering algorithms (K-means). The algorithm was implemented based on the approach mentioned in the paper and all the processes (i.e, vectorizing the reviews,

creation of tfidf matrix, assigning the cluster labels to the reviews, and others) were successfully performed. This paper does not cover the semi-supervised learning process, as most examinations reviewed do not examine classifications and explain extraction modules only.

REFERENCES

[1] Hoong-Cheng Soong, Norazira Binti A Jalil, Ramesh Kumar Ayyasamy, Rehan Akbar, "The Essential of Sentiment Analysis and Opinion Mining in Social Media", IEEE 2019.

[2] I. K. C. U. Perera and H.A. Caldera , "Aspect Based Opinion Mining on Restaurant Reviews", IEEE 2017.

[3] Casey Doyley, Alex Meandzijay, Gyorgy Kornissy, Boleslaw Szymanskiyz, Derrik Asherx, and Elizabeth Bowman , "Mining personal media thresholds for opinion dynamics and social influence", IEEE 2018.

[4] Sathis Kumar T , Mohamed Nabeem P , Manoj C K and Jeyachandran K , "Sentimental Analysis (Opinion Mining) in Social Network by Using Svm Algorithm", IEEE 2020.

[5] Aina Musdholifah and Ekki Rinaldi, "FVEC feature and Machine Learning Approach for Indonesian Opinion Mining on YouTube Comments", IEEE 2018.

[6] Malini R and Dr.Sunitha M.R, "OPINION MINING ON MOVIE REVIEWS", IEEE 2019.

[7] Untung Rahardja , Taqwa Hariguna and Wiga Maulana Baihaqi, "Opinion Mining On E-Commerce Data Using Sentiment Analysis And K-Medoid Clustering", IEEE 2019

[8] IRUM SINDHU, SHER MUHAMMAD DAUDPOTA, KAMAL BADAR, MAHEEN BAKHTYAR, JUNAID BABER and MOHAMMAD NURUNNABI, "Aspect Based Opinion Mining On Student's Feedback For Faculty Teaching Performance Evaluation", IEEE 2020

[9] Lamine FATY , Marie NDIAYE , Edouard Ngor SARR , Ousmane SALL , Sény Ndiaye MBAYE , Tony , Tona Landu , Babiga BIRREGAH and Mamadou BOUSSO, "News Comments Modeling for Opinion Mining: The Case of Senegalese Online Press", IEEE 2020

[10] Md. Mahiuddin, "Real Time Sentiment Analysis and Opinion Mining on Refugee Crisis", IEEE 2019

[11] C.N.S.Vinoth Kumar, A.Suhasini, "Secured Three-Tier Architecture for Wireless Sensor Networks Using Chaotic Neural Networks", 'Advances in Intelligent Systems and Computing' AISC Series, Springer Science + Business Media Singapore 2017 Vol. No. 507, Chapter No. 13, pp. No. 129-136,

[12] GautamSrivastava, C.N.S. Vinoth Kumar, V Kavitha, N Parthiban, RevathiVenkataraman, "Two-Stage Data Encryption using Chaotic Neural Networks", Journal of Intelligent and Fuzzy systems, Vol. no.38, Issue. No.3, pp no.2561-2568, March 2020.

[13] PraharshaSarma, Utkarsh Kumar, C.N.S.Vinoth Kumar, M.VasimBabu, "Accident Detection And Prevention Using Iot& Python Opencv", International Journal Of Scientific & Technology Research(IJSTR), Volume 9, Issue 04,pp no. 2677-2681, ISSN No: 2277-8616 April 2020.

[14] M.VasimBabu, C.N.S. Vinoth Kumar, M.Venu, International journal entitled "Improvisation of localization accuracy using ERSSI based on ADV-HOP algorithm in wireless sensor network", International journal of innovative technology and exploring engineering (IJITEE), Feb 2019

[15] A.Saranya, R.Naresh "Cloud Based Efficient Authentication for Mobile Payments using Key Distribution Method", Journal of Ambient Intelligence and Humanized Computing, Springer, 02 January, 2021. https://link.springer.com/article/10.1007%2Fs12652-020-02765-7

[16] R.Naresh, P.Vijayakumar, L. Jegatha Deborah, R. Sivakumar, "A Novel Trust Model for Secure Group Communication in Distributed Computing", Special Issue for Security and Privacy in Cloud Computing, Journal of Organizational and End User Computing, IGI Global, Vol.32, No. 3,

Septemer 2020, Pp. 1-14.

[17] R.Naresh, M.Sayeekumar, G.M.Karthick, P.Supraja, "Attribute-based hierarchical file encryption for efficient retrieval of files by DV index tree from cloud using crossover genetic algorithm",Soft Computing, Springer, Vol.23, No. 8, 2019, Pp. 2561-2574.

[18] R Divya Mounika, R.Naresh, "The concept of Privacy and Standardization of Microservice Architectures in cloud computing", European Journal of Molecular & Clinical Medicine, Vol 7, No 2, Pages 5349-5370, Dec 2020.

[19] P.Vijayakumar, R.Naresh, L. Jegatha Deborah, SK Hafizul Islam, "An efficient group key agreement protocol for secure P2P communication", Security and Communication Networks, Wiley, Vol.9, No.17, pp.3952–3965, 2016

[20] P.Vijayakumar, R.Naresh, SK Hafizul Islam, L. Jegatha Deborah "An Effective Key Distribution for Secure Internet Pay-TV using Access Key Hierarchies", Security and Communication Networks, Wiley, Vol.9, No.18, pp.5085–5097, 2016.

[21] R. Naresh, M Meenakshi, G Niranjana, "Efficient study of Smart Garbage Collection for Ecofriendly Environment", Journal of Green Engineering, Vol.10, No.1, pp.1-10,Feb 2020.