International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org
Vol. 27 No. 2 July 2021, pp.412-419

Mining Twitter Data for Business Intelligence Using Naive Bayes Algorithm for Sentiment Analysis

Ugochukwu E. Orji, Modesta E. Ezema, Jonathan C. Agbo

Department of Computer Science University of Nigeria,

Nsukka, Enugu State, Nigeria

Abstract – Today social media has grown to be a big player in the way businesses and organizations operate, especially with the coronavirus pandemic increase the online footprint of these organizations. The use of data from social media to drive business intelligence is now of growing interest to both researchers and business owners. Business owners can now utilize platforms like Twitter to learn about their target audience and improve their business processes to meet their growing needs. Twitter makes it easy to see what is going or about to go viral and vital details like why it is going viral and the players behind it. This research aims to help business owners' especial small and medium enterprises and start-ups gain a competitive advantage in their industry by using the "crowd wisdom" opportunity via social media. The proposed system is based on Twitter and crawls the platform for relevant data, including; locations, trends, and important actors (influencers) within a specified field; the system cleans the data and presents the information in an actionable format. Python was used for Twitter data mining, and sentiment analysis of the tweets was done using Naive Bayes classifiers.

Keywords – Twitter Sentiment Analysis, Twitter Sentiment Analysis for Business Intelligence, Naive Bayes algorithm, Bayes Theorem, Business intelligence, Sentiment analysis.

I. Introduction

A recent World Bank report showed that Small and Medium-scaled Enterprises (SMEs) account for 90% of businesses worldwide, especially in developing countries, making huge contributions to job creation with over 50% of employment worldwide and overall global economic development [1, 2]. The vast amount of data from social media consumption can equip organizations of different sizes to execute timely and meaningful business decisions via social media analytics [3].

Social media's usefulness as a tool for extracting patterns and behavioral traits of users cannot be underestimated, especially as it can be used to better understand the target audience and their ever-changing needs.

In the past few years, researchers from multi-disciplinary fields have been exploring intelligent ways in which the feedback from social media platforms like Twitter can feed actionable information to businesses and organizations for business intelligence.

However, the major challenge has been the inherent noise or the unstructured nature of social media data [4]. No doubt, the dynamic nature of such data and its massive size pose a significant challenge for researchers, but with modern machine learning approaches to data mining and analytics, it is now possible to minimize errors and sieve out useful information from the 'wisdom of the crowd' available on Twitter and other social media platforms [5].

Corresponding Author: Ugochukwu E. Orji

II. Theoretical background

Business intelligence explores how to comprehensively and scientifically tease out vital answers to questions bugging businesses and organizations, which can then be organized and stored in a result-oriented way to provide business value and improved decision-making [6, 7].

When business intelligence tools and techniques are applied to Twitter data, SMEs can gain a competitive edge, cut costs, and release products with a higher degree of success. It also makes advertising and marketing of goods and services easier and well-targeted for success [8].

Social media usage is currently one of the most popular online activities and on the steady rise. A Sprout Social report estimates that in 2020, over 3.6 billion people all over the world use one or two social media platforms, a number which they also projected to increase to almost 4.41 billion by 2025 [9]. Another Sprout Social report showed that in 2020, global retail e-commerce sales peaked at 4.28 trillion U.S. dollars and are projected to reach 5.4 trillion U.S. dollars in 2022 [10].

There is also reciprocal growth in the power social media has in users' purchasing awareness and decisions. According to bigcommerce, 43% of global shoppers research products online via social networks before purchasing [11].

Furthermore, [12] reports that social media is the number one influencer of fashion shopping amongst 18-24 years women. The report shows that 32% of women aged 18-34 are open to buying all their fashion needs directly from a social media platform.

This also demonstrates the power social media has and, if appropriately harnessed, could give SMEs a good competitive advantage.

2.1 Concept of Twitter Data Mining

With data mining tools and technology, it is now possible to discover useful and actionable knowledge in large-scale data sources like Twitter and other social media platforms. Data mining is an integral part of many related fields, including statistics, machine learning, pattern recognition, database systems, visualization, data warehouse, and information retrieval [13, 14].

Mining Twitter data involves the process of extracting actionable patterns from the Twitter data pool that can be beneficial for businesses, users, and other consumers. SMEs and start-ups can get customer feedback for proactive planning, develop recommendation systems to maintain existing customers and hopefully even gain new customers via harnessing data from Twitter [15].

2.1.1 Common issues encountered in Mining Twitter data

Data generated on social media sites are unstructured (not organized in a pre-defined manner) and different from the conventional structured data sources [16]. Social media data are mostly user-generated content; thus, social media data are vast, noisy, distributed, unstructured, and dynamic.

The implication for data mining tasks is that new efficient techniques and algorithms are required to make sense of social media data [17]. For example, according to Twitter's Q3 2019 report, the platform has 145 million monetizable daily active users [18], while Facebook had over 2.6 billion monthly active users in the first quarter of 2020 [19]. YouTube reported over 2 Billion logged-in users' visits each month [20], and Wikipedia hosts over 54 million articles attracting an estimated 1.5 billion unique visitors per month [21]. This just shows the abundance of potentials available on social media and the difficulty in making sense of the data.

The following are the major issues associated with mining social media data:

- Social media data are noisy: removing the noise from social media data is paramount for effective mining. Research has shown that spammers generate more data than legitimate users [22, 23]. Hence, separating legitimate users from bots, trolls, and scammers becomes necessary while mining social media data.
- Social media data are distributed: because there is no central authority that maintains data from all social media sites, it poses a daunting task for researchers trying to understand the information flow.
- Social media data are often unstructured: it will always be a big challenge trying to make meaningful observations based on unstructured data from various data sources. For example, LinkedIn, Facebook, and Instagram serve different purposes and meet the diverse needs of their users respectively.

- Social media sites are dynamic and continuously evolving: The dynamic nature of social media data is a significant challenge for researchers, especially the speed with which social media sites evolve.

2.2 Sentiment Analysis techniques

Sentiment analysis is also known as opinion mining, and it is used to extract opinions expressed in user-generated content like tweets [24]. Sentiment analysis tools help businesses and organizations to understand product/service sentiments, brand perception, new product perception, and reputation management [25].

2.2.1 Steps to conduct sentiment analysis

Major steps of conducting sentiment analysis include;

- Finding relevant documents,
- Finding relevant sections,
- Finding the overall sentiment,
- Quantifying the sentiment, and
- Combining all sentiments for better understanding [26].

Basic components of opinion are;

a) an object on which opinion is expressed,

b) an opinion expressed on an object and;

c) the opinion holder.

Here, objects represent a finite set of features, each symbolizing a set of synonymous words or phrases. Opinion mining tasks are performed at the document, sentence, or feature level [27, 28, 29].

2.3 Machine learning approach to Sentiment Analysis

To implement sentiment analysis on text classification problems, machine learning algorithms are used with the help of feature selection techniques which selects only important features by eliminating the noisy and irrelevant features found in the dataset [30].

2.3.1 Using Naïve Bayes (NB) Classifier for Sentiment Analysis

Naïve Bayes classifier is one of the most commonly used classifiers. This classification algorithm uses Bayes Theorem to predict the probability of a given feature set belonging to a particular model [31]. Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities, which measures the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred [32]. It works by computing the posterior probability of a class; its computation is hinged on the distribution of words in the document. The Naïve Bayes Classification model works with the Bag of words (BoW) feature extraction. The BoW feature extraction does not take into cognizance the position of a word in the document [33]. One significant advantage of the Naïve Bayes (N.B.) classifier is that it has a fast execution time and low memory consumption [34].

For an example of a classification problem based on the Naive Bayes algorithm see below. To find the probability for a label, we use the Bayes rule to express it as:

P(label|features) in terms of P(label) and P(features|label):

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features}|\text{label})}{P(\text{features})}$$

Where: P(label) defines the prior probability of a label, P(features|Label) defines the prior probability that a given feature set is classified as a label, P(features) defines the prior probability that a given feature set is occurred.

The algorithm then makes the 'naive' assumption that all features are independent, given the label:

$$P(label|features) = \frac{P(label) * P(f1|label) * ... * P(fn|label)}{P(features)}$$

Rather than computing P(features) explicitly, the algorithm just calculates the numerator for each label, and normalizes them so they sum to one:

$$P(label|features) = \frac{P(label) * P(f1|label) * ... * P(fn|label)}{SUM[l](P(l) * P(f1|l) * ... * P(fn|l))}$$

2.4 Related works

This study briefly sampled the research from other researchers in mining social media data, sentiment analysis, and business intelligence.

In [35], the researchers applied two proposed models of classification, Naive Bayes algorithm and G.A., as feature selection into digital learning application review data.

In [36], the authors applied sentiment analysis to find customer satisfaction levels from digital payment services in Indonesia with percentage accuracy and conclusion opinion mining.

The authors in [37] analyzed how the citizens of different countries are dealing with the COVID 19 situation. Tweets were collected, pre-processed, and then used for text mining and sentiment analysis. The results showed how people around the world were reacting to the pandemic based on their Tweets.

Likewise, in [38], the authors studied the use of Twitter data for waste minimization in the beef supply chain. They used Twitter Application Programming Interface (API) to extract Twitter data. Then they did a combination of Sentiment Analysis, Descriptive Analysis, and Content Analysis of the tweets and sorted the data accordingly. The authors then linked the individual complaints to known roots causes in the beef supply chain. The information helped mitigate food waste and improve customer satisfaction.

In [5], the authors proposed CNN-BiLSTM- and BERT-based deep neural models to combine attitude representation and content representation for early rumour detection on Twitter. They experimented on real-world rumour datasets using the BERT-based model to achieve effective early rumour detection on Twitter.

The author in [39] applied sentiment analysis with the Trip advisor dataset, a travel website that provides user opinions about hotels and restaurants. The analysis task categorized into three steps, polarity detection, Aspect selection, and classification. The author applied SAMs, SentiStrength, Bing, Syuzhet, and CoreNLP methods to analyze the user opinions.

In this study [40], the author proposed sentiment analysis on electronic products like mobile devices and laptops using machine learning algorithms. They presented a new vector feature for classifying the reviews into positive and negative. User reviews about electronic products taken from Twitter as input dataset and performed sentence-level sentiment analysis. Naive Bayes, SVM, Maximum Entropy Classifier and Ensemble classifiers are applied.

III. ALGORITHM DESIGN OF THE SENTIMENT ANALYSIS ON TWITTER DATA

The following explains the process of getting the sentiments on the tweets:

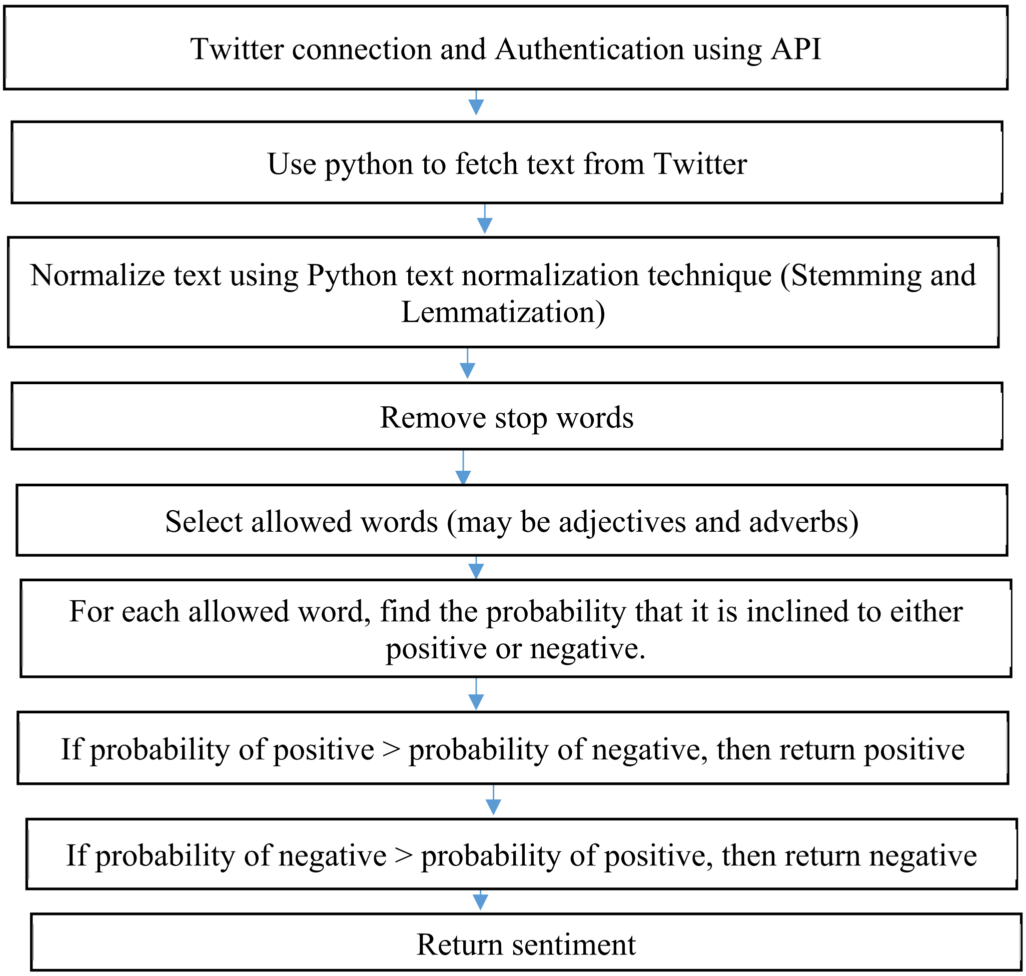

Fig.1.Flowchart of sentiment analysis on Tweets

IV. DATA GATHERING AND PROCESSING

The IMDB dataset is an already trained open-source dataset freely available on the internet for this particular purpose and was modeled in this project. The data sets contain 10,000 movie reviews and available at: https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

The data was divided into 60% training set and 40% test set and stored the dataset in the database. The sentiments for the tweets were done as illustrated in fig.1.0 above.

Overall, 1,000 tweets were collected from each entered keyword after every five days. The keywords used for collecting the tweet were; DOGECOIN, IPHONE 12, HUAWEI P40 and HP FOLIO.

The data was cleaned and prepared for analysis by filtering the retweets and replies to avoid duplicating the tweets. After collecting the data to the database, the data normalization process was done, where the white spaces, punctuation, stop words were removed, and the tweets were converted to lower case. After the data cleaning, naïve bayes algorithm through python was performed on the tweets. Once the scoring of the tweets was done, it was saved.

V. RESULTS AND DISCUSSION

The platform was able to generate, and process sentiment analysis on the different keywords crawled from Twitter. The result was sorted by datestamp, keyword location and sentiment ratio for each tweet fetched. The chart of the results is shown in fig. 2 below.

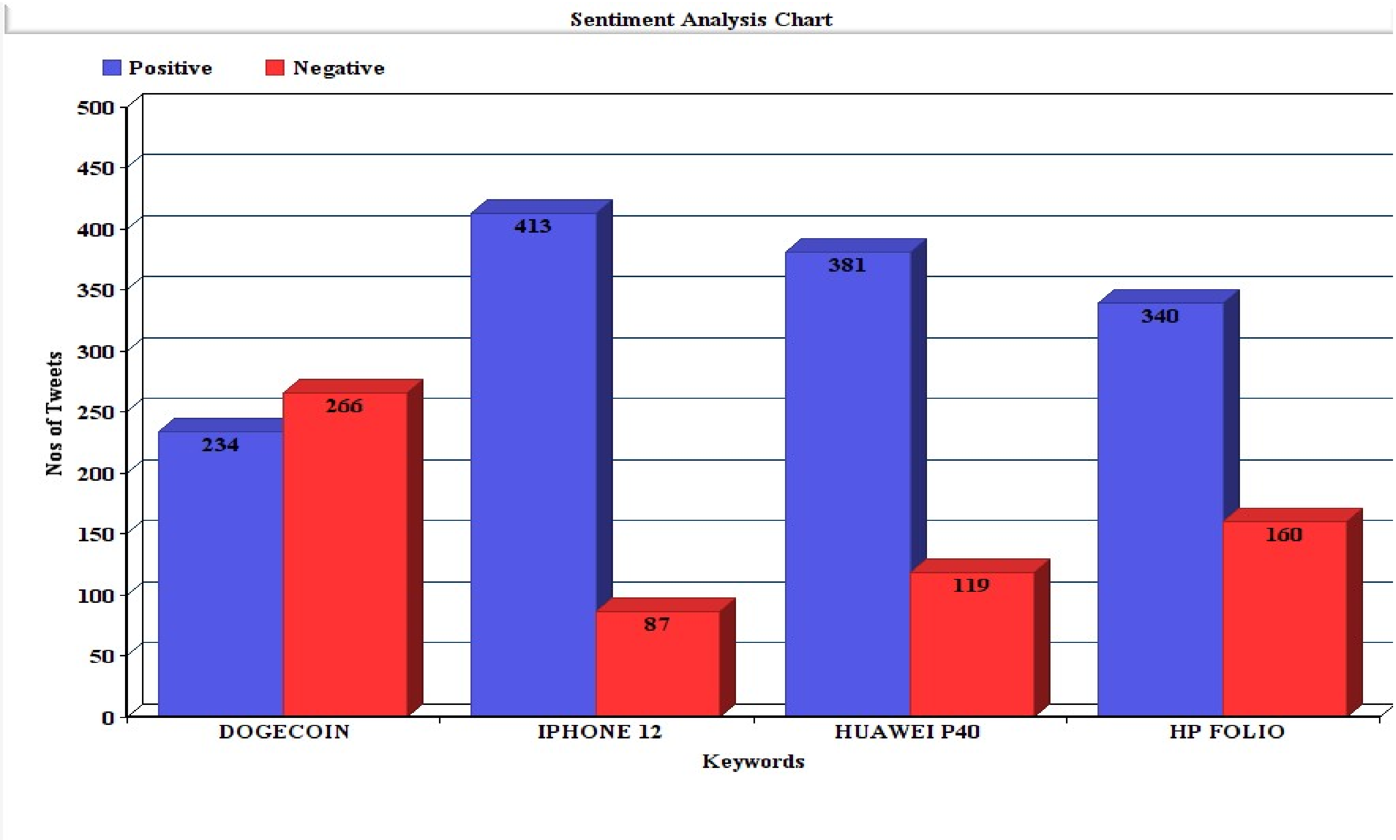

Fig. 2 Result chart of Sentiment Analysis of Tweets

With this tool, SMEs can stay updated with their potential buyers' needs by harnessing Twitter data; they can also see user feedback on products in the market by getting sentiment analysis on those products before stocking up. This ensures increased turn up and overall sustainability of their business.

VI. Conclusion, Limitation and Future Scope

This research work aimed at analyzing the sentiments from Twitter data for business intelligence. The study shows how Twitter users react to different selected keywords. This data-driven tool could potentially be a big deal for SMEs to improve their business strategies and decision-making.

There is always room for improvement. This research targeted only Twitter for its data, but there are so many other social media platforms like Facebook, Instagram, etc., where more data can be harnessed to better understand buyers.

The tweets collected for this study were in English, which might serve as a limitation for the study. Also, the python code used for this study does not count emoticons (emojis) which is another limitation.

For future works, this study can analyze the changing emotions and sentiments in various fields of life and could be a very handy tool for start-ups and SMEs.

References

[1] Banji Oyelaran-Oyeyinka., "SME: Issues, Challenges and Prospects," FSS 2020 International Conference.

[2] “Nigeria SME survey: Assessing current market conditions and business growth prospects.” Accessed on: March 23, 2021. [Online] Available at: https://www.pwc.com/ng/en/events/nigeria-sme-survey.html

[3] U. Sivarajah, Z. Irani, S. Gupta, and K. Mahroof; "Role of big data and social media analytics for business to business sustainability: A participatory web context." Industrial Marketing Management 86 (2020): 163-179.

[4] R.A. Callcut, S. Moore, G. Wakam, A.E. Hubbard, and M.J. Cohen; "Finding the signal in the noise: Could social media be utilized for early hospital notification of multiple casualty events?." Plos one 12, no. 10 (2017): e0186118.

[5] Tian L., Zhang X., Wang Y., Liu H. (2020) Early Detection of Rumours on Twitter via Stance Transfer Learning. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12035. Springer, Cham. https://doi.org/10.1007/978-3-030-45439-5_38

[6] E. Foley, & M. G. Guillemette, "What is business intelligence?" International Journal of Business Intelligence Research, Vol. 1(4), pp. 1–28, 2010.

[7] D. J. Power, & G. Phillips-Wren, "Impact of social media and Web 2.0 on decision-making." Journal of decision systems, Vol. 20, pp. 249–261, 2011.

[8] Vincent Dutot & Elaine Mosconi (Guest Editors); "Social media and business intelligence: defining and understanding social media intelligence," Journal of Decision Systems, Vol. 25:3, pp. 191-192, 2016.

[9] “Sprout Social report on Number of global social network users 2017-2025;” Accessed on: June 23, 2021. [Online] Available at: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[10] “Global retail e-commerce sales 2014-2024;” [Online] Available at: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

[11] “16 Online Shopping Statistics: How Many People Shop Online?” [Online] Available at: https://www.bigcommerce.com/blog/online-shopping-statistics/#5-essential-online-shopping-statistics

[12] “Mintel Report on Clothing retailing;” Accessed on: March 23, 2021. [Online] Available at: http://reports.mintel.com/display/859493/#

[13], P. N. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining”. Pearson Addison Wesley: Boston, 2006.

[14] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques”, Morgan Kaufmann: San Francisco, 2011.

[15] V. Tundjungsari; “Business Intelligence with Social Media and Data Mining to Support Customer Satisfaction in Telecommunication Industry.” International Journal of Computer Science and Electronics Engineering (IJCSEE) Volume. 2013;1.

[16] Pritam GundechaHuan Liu. "Mining Social Media: A Brief Introduction." In INFORMS Tutorials in Operations Research. Published online: 14 Oct 2014; pp. 1-17. https://pubsonline .informs.org/doi/pdf/10.1287/educ.1120.0105 [Accessed 20 Nov 2019].

[17] G. Bello-Orgaz, J. J. Jung, D. Camacho; “Social big data: Recent achievements and new challenges.” Information Fusion. Vol. 1;28:45-59. Mar. 2016

[18] Twitter Q3 2019 Letter to Shareholders; https://s22.q4cdn.com/826641620/files/ doc_financials/2019/q3/Q3-2019-Shareholder-Letter.pdf (accessed May, 2020)

[19] Number of monthly active Facebook users worldwide as of 2nd quarter 2020; https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (accessed May, 2020)

[20] YouTube for Press; https://www.youtube.com/about/press/ (accessed May, 2020).

[21] From Wikipedia, the free encyclopedia; https://en.wikipedia.org/wiki/Wikipedia (accessed May, 2020).

[22] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. "Detecting spam in a Twitter network." First Monday 15.1 (2010).

[23] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. "Who is tweeting on Twitter: Human, bot, or cyborg?" Proceedings of the 26th Annual Computer Security Applications Conference. Association for Computing Machinery, New York, pp. 21-30, 2010.

[24] The Free Library. S.V. “A comparative study on sentiment analysis.” Accessed on: March 23, 2021. [Online] Available at: https://www.thefreelibrary.com/A+comparative+study+on+sentiment+analysis.-a0385070617

[25] B. Pang and L. Lee. "Opinion mining and sentiment analysis." Foundations and Trend in Information Retrieval Vol. 2.1–2, pp. 1-135, 2008.

[26] A. M. Popescu and O. Etzioni. "Extracting product features and opinions from reviews." Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, pp. 339-346, 2005.

[27] H. Liu and P. Maes. “InterestMap: Harvesting social network roles for recommendations.” Workshop: Beyond Personalization, San Diego, 2005.

[28] E. Rilo and J. Wiebe. "Learning extraction patterns for subjective expressions." Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, pp. 105-112, 2003.

[29] H. Yu and V. Hatzivassiloglou. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences & Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, pp. 129-136, 2003.

[30] B. Agarwal, N. Mittal; “Machine Learning Approach for Sentiment Analysis. In: Prominent Feature Extraction for Sentiment Analysis.” Socio-Affective Computing. Springer, Cham. 2016 https://doi.org/10.1007/978-3-319-25343-5_3

[30] 31S.S. Kamble, A.R. Itkikar; “Study of supervised machine learning approaches for sentiment analysis.” International Research Journal of Engineering and Technology (IRJET). Vol. 5(04). Apr. 2018.

[32] Nagesh Singh Chauhan; “Naïve Bayes Algorithm: Everything you need to know.” Accessed on: March 23, 2021. [Online] Available at: https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

[33] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, & M. Demirbas; “Short text classification in twitter to improve information filtering.” In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval 2010 Jul. 19 (pp. 841-842).

[34] K.W. Kiprono, E.O. Abade; “Comparative Twitter sentiment analysis based on linear and probabilistic models.” International Journal on Data Science and Technology. 2016;2(4):41-5.

[35] Ernawati, Siti, Risa Wati, Nuzuliarini Nuris, Lita Sari Marita, and Eka Rini Yulia. "Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application." In Journal of Physics: Conference Series, vol. 1641, no. 1, p. 012040. IOP Publishing, 2020.

[36] Hilman Wisnu et al, Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes 2020 J. Phys.: Conf. Ser. 1444 012034

[37] Dubey Akash Dutt, “Twitter Sentiment Analysis during COVID-19 Outbreak” Accessed on: March 23, 2021. [Online] Available at: SSRN: https://ssrn.com/abstract=3572023 or http://dx.doi.org/10.2139/ssrn.3572023

[38] N. Mishra, A. Singh; “Use of twitter data for waste minimization in beef supply chain.” Ann Oper Res 270, 337–359 (2018). https://doi.org/10.1007/s10479-016-2303-4

[39] Ana Valdivia M. Victoria Luzón, and Francisco Herrera; “Sentiment Analysis in TripAdvisor;” University of Granada, IEEE INTELLIGENT SYSTEMS, Published by the IEEEComputer Society.

[40] Neethu M S,Rajasree R, “Sentiment Analysis in Twitter using Machine Learning Techniques.;” IEEE - 31661,4th ICCCNT 2013 July 4 - 6, 2013,