2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON) | 978-1-6654-7978-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/NIGERCON54645.2022.9803087

Using Twitter Sentiment Analysis for Sustainable Improvement of Business Intelligence in Nigerian Small and Medium-Scale Enterprises

Ugochukwu .E. Orji
Dept. of Computer Science
University of Nigeria, Nsukka,
Enugu, Nigeria.
ugochukwu.orji.pg00609@unn.edu.ng

Modesta .E. Ezema
Dept. of Computer Science
University of Nigeria, Nsukka,
Enugu, Nigeria.
modesta.ezema@unn.edu.ng

Jideofor Ujah
Dept. of Computer Science
University of Nigeria, Nsukka,
Enugu, Nigeria.
jideofor.ujah@gmail.com

Ponsak .S. Bande
Dept. of Computer Science
University of Nigeria, Nsukka,
Enugu, Nigeria.
ponsak.bande@unn.edu.ng

Jonathan C. Agbo
Dept. of Computer Science
University of Nigeria, Nsukka,
Enugu, Nigeria.
agbo.jonathan@unn.edu.ng

***Abstract* — For most Small and Medium-scale Enterprises (SMEs) and startups in Nigeria, conducting market research and evaluation is solely based on customers' verbal feedback and demands, which is often flawed and highly limited. Social media (Twitter), as a business intelligence tool, can help SMEs gain insight into customers' perspectives of their products and services through Sentiment Analysis. This research presents a Twitter Sentiment Analysis for business intelligence using three machine learning algorithms; Bernoulli Naive Bayes (BNB), Linear Support Vector Classification (SVC), and Logistic Regression (LR) algorithm, to get the text polarity of tweets (Negative/positive). The dataset used for the research was gotten from Kaggle; the tweets were processed and analyzed using Python programming libraries on Kaggle's Jupyter Notebook cloud environment. Our result showed that Logistic Regression achieved better performance with an accuracy of 83% based on Precision-Recall evaluation metrics.**

Keywords— SMEs, Business Intelligence, Twitter Sentiment Analysis, Machine Learning Algorithm.

I. Introduction

The Central Bank of Nigeria (CBN) defines Small and Medium-scale Enterprises (SMEs) as businesses with a turnover of less than ₦500,000,000 (excluding land and buildings) per annum and/or employs less than 300 people. SMEs account for about 96% of businesses and 84% of employment in Nigeria [1] [2]. One could deduce that the profitability and growth of SMEs would equally mean positive growth for the Nigerian economy.

The use of social media has transformed business-customer relationships and inspired more businesses to seek information from the "wisdom of the crowd." According to [3], data from User-Generated Content (UGC) can feed actionable information to organizations of all sizes, as they seek to make timely and profitable business decisions. The main objective of Social Media Analytics (SMA), which includes Twitter Sentiment Analysis (TSA), is to create tools and techniques for evaluating, collecting, visualizing, and monitoring data from social media [4]. The feedback from TSA is key to effective business intelligence because it helps business owners and other users develop a corresponding decision-making framework and solution to help their business grow [4]. Most SMEs in Nigeria still depend on the conventional market survey method; based on routine customer demands. However, in today's fast-paced society, it is completely naive not to take advantage of the possibilities available through social media for business growth. As a startup or SME, it is unwise to blindly enter the market without adequately exploring the avenues available to you. Hence, Nigerian SMEs must be ready to explore data-driven approaches to their business to avoid losing money and customers to better-prepared competitors. The use of TSA for business intelligence is highly recommended because, at the moment, everything from buying a pair of shoes to ordering food or booking flights can be done on a mobile device; this shows that more businesses are now done online. A recent study by [5] estimates that over 1.7 billion people now have accounts on one or more social media platforms like Twitter which is helping to bridge the gap and connecting people from all over the world.

Thus, this project seeks to help Nigerian SMEs enhance their productivity and efficiency by harnessing the power of social media analytics for business intelligence. Currently, there is limited research on the actual value of social media for SMEs. This research intends to fill this gap in literature by exploring ways SMEs can collect, analyze, and use information from social media to grow and strengthen their businesses.

II. Literature review

The following section explains the key concepts covered in this research and the review of related literature on the concepts and methods.

A. *Key Concepts*

1) Business intelligence (B.I): The authors in [6] credit Hans Peter Luhn with the term Business Intelligence System. Luhn's definition of business intelligence is the summation of all the processes, architectures, and

978-1-6654-7978-3/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: University of Arkansas Litte Rock. Downloaded on March 07,2023 at 04:31:01 UTC from IEEE Xplore. Restrictions apply.

technologies used to transform raw and unstructured data into business-oriented information. Steps to B.I include:

- Gathering and accessing data.Data integration, cleaning, and preparation.
- Data mining and analysis using machine learning algorithms and tools.
- Creating visualization/ Result presentation [7].

2) Sentiment Analysis: This falls under the Natural Language Processing (NLP) technique of artificial intelligence that extracts opinions/expressions from UGC like social media posts. Sentiments in UGC are usually categorized as either positive, negative, or neutral. Here is an example:

- “I like the new look of your living room!” → Positive.
- “I’m unsure about the new design of your living room.” → Neutral.
- “The new design of your living room is awful!” → Negative

Twitter Sentiment Analysis (TSA) is a specific aspect of sentiment analysis that deals with mining opinions from UGC on Twitter. Approaches devised for TSA are the same as those developed for conventional NLP in other areas and are valuable for extracting vital data like product reviews [8]. You can perform opinion mining tasks at various levels, including; the document, sentence, and feature level [9][10].

B. *Review of Related Literature*

TSA has proven to be a vital tool for monitoring business and social trends on the Twitter platform. TSA can predict user perception of a company’s product or service, which aids decision-making. The importance of retrieving accurate information through TSA cannot be overstated, as demonstrated by the authors in [11]. The study evaluated 28 top academic and commercial systems with the objective of assessing the recent innovations in TSA. The authors looked at Twitter sentiment classification applied to five different Twitter-related datasets and the result showed that most available systems differed across domains. They also went further to rank each system according to its effectiveness.

The authors in [12] carried out a strategic overview of the algorithms and current approaches utilized for TSA in multiple fields such as health, stock sales, disaster management, etc. The authors reviewed over 40 TSA-related articles and categorized them into approaches that integrate machine learning and lexicon-based techniques with theories and technologies from cognitive science, semantic web, and big data. The paper showed TSA’s importance in many active and future academic and industry research.

The authors in [13] highlighted the practical importance of Business Intelligence for small-sized companies. The authors focused on two critical areas of every business: profitability and customer satisfaction. They interviews 50 entrepreneurs using a questionnaire to tease out the entrepreneur’s competencies, customer relationship management (CRM), and business intelligence approaches. The survey result showed that most small companies do not use data-driven approaches for their business.

TSA has become an important medium for gaining insight into people’s views on topics, products, or services. In [14], the author used TSA to show how the people of Nepal were reacting to the global COVID-19 outbreak. The author collected data for the research using the Tweepy Python library and then used the TextBlob library to get the polarity of the tweets (positive, negative, or neutral text).

Finally, in a related work done by authors in [15], three machine learning classifiers, such as the Naive Bayes classifier, Maximum Entropy (MaxEnt) model, and the Support Vector Machines (SVM) were deployed to classify the sentiment of tweets using the distant supervision approach. They trained their model with emoticon data and achieved 80% performance accuracy. Our approach achieved better performance accuracy because we introduced a better technique for our data processing.

III. Methodology

We deployed three machine learning algorithms: BNB, SVC, and LR algorithms, to get the sentiment polarity of tweets in the dataset and classify the tweets into negative/positive/neutral sentiments. The idea is to show how SMEs and other start-ups can use Twitter’s “Wisdom of the crowd” capacity to track how users and customers react to their products and services.

Fig. 1 below describes the steps taken to achieve our results in this study.

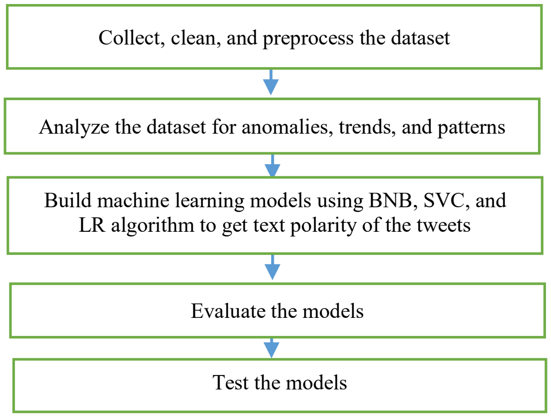

Collect, clean, and preprocess the dataset
Analyze the dataset for anomalies, trends, and patterns
Build machine learning models using BNB, SVC, and LR algorithm to get text polarity of the tweets
Evaluate the models
Test the models

Fig. 1: Research steps

1.Dataset: The dataset utilized in this research is the “sentiment140 dataset.” The dataset contains 1,600,000 tweets freely available on Kaggle [16] and is licensed under the Creative Commons designation (CC0 1.0). Table 1 below gives a brief description of the dataset attributes.

Authorized licensed use limited to: University of Arkansas Litte Rock. Downloaded on March 07,2023 at 04:31:01 UTC from IEEE Xplore. Restrictions apply.

Table 1: dataset description

Variable Name	Description	Data Type
I.D	Tweet I.D	Numeric
User	Username of tweeps	String
Text	The text of the tweet	String
Date	Tweet date	Datestamp
Sentiment	The classification of the tweets (0 = negative, 1 = positive)	Categorical

2. Data cleaning and preprocessing

The data preprocessing steps taken in this study include:

a. TF-IDF Vectoriser: To process natural language data, the text must be converted and represented as a numerical feature. This process of transforming text into a numerical

  a. feature for effective modeling is called text vectorization. TF-IDF is one of the most popular text vectorizers [17] [18]. It is given as follows;

$$w_{i,j} = tf_{i,j} \times idf_i \qquad [19]$$

Where;

$w_{i,j}$ is the vectorized score for i in j; $tf_{i,j}$ is the term frequency for i in j; and idf_i is the vectorized score for i.

b. Lemmatization is used to convert a word to its base form (e.g., “Walking” to “Walk”)

c. Converted each text to lowercase.

  a. Standarded the URLs: This entails changing it from a variety of "http", "https" & "www" to just the "URL" value.

b. Created a pre-defined emoji dictionary alternative to replace the unstructured ones in the UGC.

c. Standardized the Usernames.

d. Removed Non-Alphabets: Only retaining digits, alphabets, and space.

e. Shortened or removed where there are three (3) or more consecutive letters.

f. Removed stopwords like “a,” “is,” and “are” to reduce noise.

3. Data Analysis

Exploratory Data Analysis (EDA) was carried out to quickly explore the dataset and uncover any implicit patterns, spot anomalies, and check assumptions. EDA is achieved via the help of summary statistics functions and graphical representations.

IV. Model Development and Result

All three models deployed in this research were implemented on Kaggle’s Jupyter Notebook cloud environment. For this, we used python version 3.7.12. The libraries used to train and test the models include, Numpy, Pandas, Scikit-learn, NLTK, Matplotlib, and Seaborn libraries. The dataset was divided into a ratio of 80:20 for training and test sets, respectively.

The results were evaluated through a confusion matrix and a Scikit library classification report of precision, recall, and F1-score. Evaluation metrics such as the precision-recall used in this research explains how well the models perform when tested. The formula is given as follows;

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive } + \text{false positive}} \text{ and Recall} = \frac{\text{True positive}}{\text{True positive} + \text{false negative}} \quad [20]$$

Where;

i. “True positive” represents correct actual positive predictions,

ii. “False positive” represents incorrect actual positive predictions,

iii. “True negative” represents correct actual negative predictions,

iv. “False negative” represents incorrect actual negative predictions and,

The “Accuracy” is the total of all correct predictions by the model.

a. Bernoulli Naive Bayes (BNB) Model

BNB is a classification algorithm that falls under supervised machine learning. The BNB model uses the principles of the Bayes theorem, which calculates the conditional probability that an event (A) will occur based on prior knowledge of another related event (B). The formula for the Bayes theorem is given as follows:

$$P(A \mid B) = \frac{P(B|A).P(A)}{P(B)} \qquad [21]$$

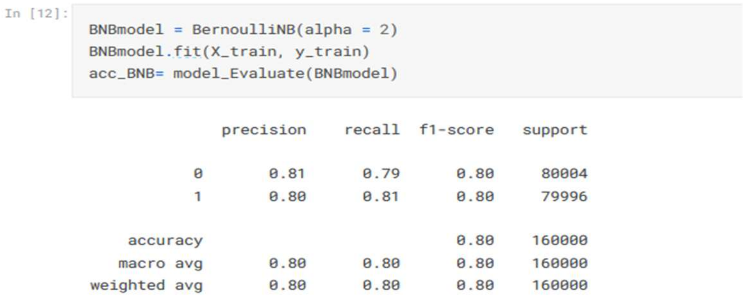
Where,

$P(A \mid B)$ = Probability of A where B holds true

$P(B \mid A)$ = Probability of B where A holds true

$P(A)$, $P(B)$ = The probability that A and B occurs

Fig. 2 below shows the evaluation result of our BNB model.



```
In [12]:
BNBmodel = BernoulliNB(alpha = 2)
BNBmodel.fit(X_train, y_train)
acc_BNB= model_Evaluate(BNBmodel)
```

```
              precision    recall  f1-score   support

           0       0.81      0.79      0.80     80004
           1       0.80      0.81      0.80     79996

    accuracy                           0.80    160000
   macro avg       0.80      0.80      0.80    160000
weighted avg       0.80      0.80      0.80    160000
```

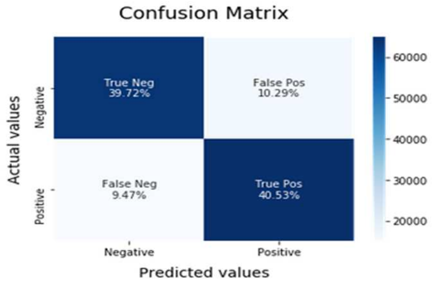


Fig. 2: Confusion matrix for BNB model

Authorized licensed use limited to: University of Arkansas Litte Rock. Downloaded on March 07,2023 at 04:31:01 UTC from IEEE Xplore. Restrictions apply.

b. Linear Support Vector Classification (SVC) Model

SVC is a supervised learning algorithm developed for linear SVM kernel computations. It is very effective for classification problems and is renowned for achieving higher speed and better performance on limited sample datasets. SVC works by;

i. Mapping data points to a high-dimensional feature space and,

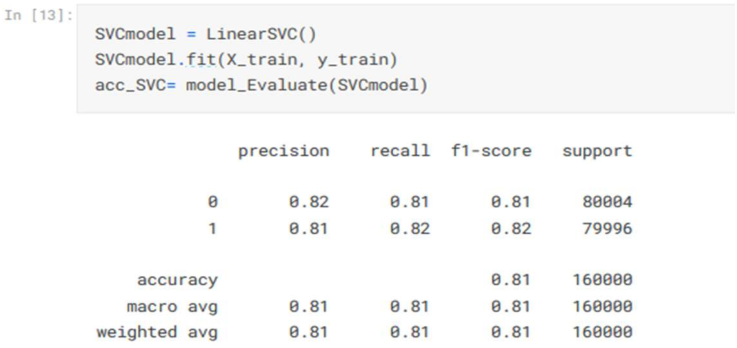
ii. Finding a separator

The formula for linear SVM kernel is given as follows:

$$f_{(x)} = \beta_{(0)} + \Sigma(a_i \; * (x, x_i)) \qquad [22]$$

Where; x_i is the support vector; $\beta_{(0)}$ and a_i are the coefficients (for each input), and x represents new input vectors.
SVC applications include; image classification, text mining, sentiment analysis, regression, and many more.

Fig. 3 below shows the evaluation result of our SVC model.



```
In [13]:
    SVCmodel = LinearSVC()
    SVCmodel.fit(X_train, y_train)
    acc_SVC= model_Evaluate(SVCmodel)
```

```
              precision    recall  f1-score   support

           0       0.82      0.81      0.81     80004
           1       0.81      0.82      0.82     79996

    accuracy                           0.81    160000
   macro avg       0.81      0.81      0.81    160000
weighted avg       0.81      0.81      0.81    160000
```

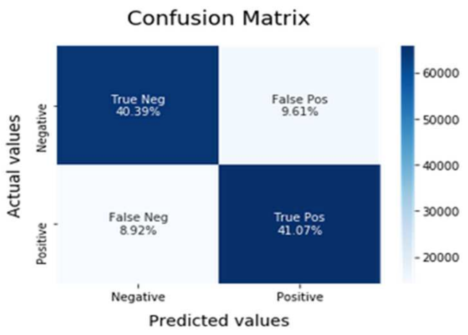


Fig. 3: Confusion matrix for SVC model

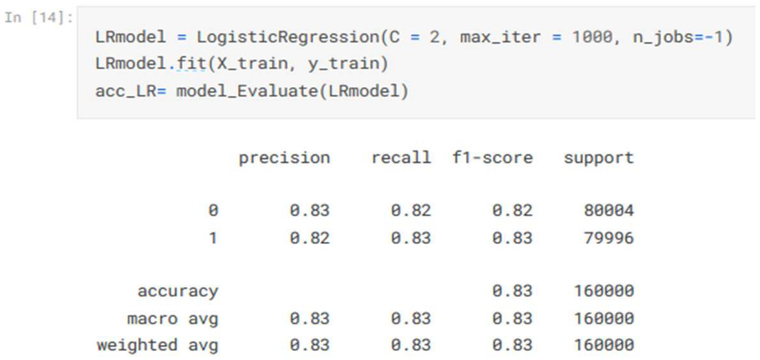
c. Logistic Regression (LR) Model

LR is a classification algorithm that is most suitable for categorizing binary data (yes/no). The equation for logistic regression is given as follows;

$$\ell = \text{logb}\left[\frac{y}{1-y}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n \qquad [23]$$

Where ℓ is the log-odds, b is the base of the logarithm, and β are the model's parameters.

Fig. 4 shows the evaluation result of our LR model.



```
In [14]:
    LRmodel = LogisticRegression(C = 2, max_iter = 1000, n_jobs=-1)
    LRmodel.fit(X_train, y_train)
    acc_LR= model_Evaluate(LRmodel)
```

```
              precision    recall  f1-score   support

           0       0.83      0.82      0.82     80004
           1       0.82      0.83      0.83     79996

    accuracy                           0.83    160000
   macro avg       0.83      0.83      0.83    160000
weighted avg       0.83      0.83      0.83    160000
```

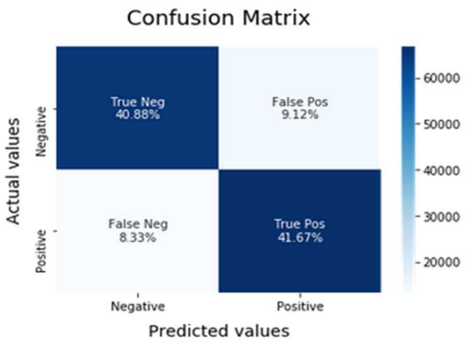


Fig. 4: Confusion matrix for LR model

V. Discussion

The performance accuracy achieved can be traced to the preprocessing techniques deployed for the research, like the TF-IDF-Vectoriser, which proved to be instrumental in ensuring that our dataset is well-processed and our models produced good accuracy. All our models for this research achieve an f1-score of 80% and above, which can be considered successful. As demonstrated in Fig. 5, the inference from the models showed that LR and the SVC models had better performance accuracy than the BNB model.

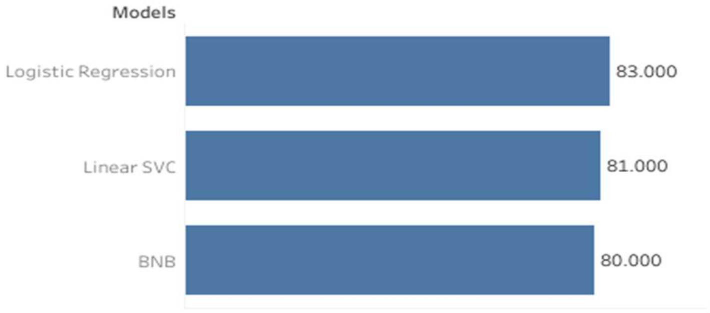
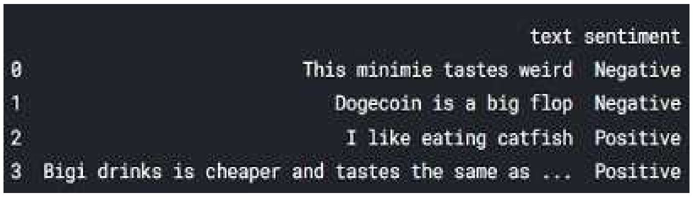


Fig. 5: Model Result Comparison

Authorized licensed use limited to: University of Arkansas Litte Rock. Downloaded on March 07,2023 at 04:31:01 UTC from IEEE Xplore. Restrictions apply.

Furthermore, the model showed great promise when tested with real-time data, as demonstrated in Fig. 6 below.



```
                                           text sentiment
0                         This minimie tastes weird  Negative
1                            Dogecoin is a big flop  Negative
2                             I like eating catfish  Positive
3  Bigi drinks is cheaper and tastes the same as ...  Positive
```

Fig. 6: Model testing

VI. CONCLUSION

This research has provided a new approach for SMEs to improve their business strategies using the “wisdom of the crowd” feature of UGC from Twitter. Social media platforms like Twitter present startups and SMEs with an avenue to directly get feedback on their products and services from their target audience. We deployed three machine learning algorithms (/BNB, SVC, and LR) to perform TSA classification and showed how SMEs could use TSA to gain insight into customer perception of their products/services. Our method also contributes to literature on how data can be collected, cleaned, and preprocessed for TSA. In conclusion, even though business intelligence application for SMEs is still at an early stage, this research’s methodological and empirical contributions could help SMEs in Nigeria make a paradigm shift to become more data-driven. Finally, this research will be especially beneficial to SMEs that has online presence on Twitter and are looking to take advantage of the opportunities available to them.

ADDITIONAL INFORMATION

The datasets analyzed, complete documentation of the data analysis, and model development process are available at: https://www.kaggle.com/orjiugochukwu/twitter-sentiment-analysis

REFERENCES

[1] O. O. Banji, “SME: Issues, Challenges, and Prospects.” Presentation at FSS 2020 international conference. 2010.

[2] PricewaterhouseCoopers, “Nigeria SME Survey,” PwC. [Online]. Available: https://www.pwc.com/ng/en/events/nigeria-sme-survey.html. [Accessed: 18-Dec-2021].

[3] U. E. Orji, M. E. Ezema, and J. C. Agbo, “Mining Twitter Data for Business Intelligence Using Naive Bayes Algorithm for Sentiment Analysis.” International Journal of Progressive Sciences and Technologies, 2021, Vol. 27(2), pp. 412-419.

[4] D. Zeng, H. Chen, R. Lusch, and S. H. Li, “Social media analytics and intelligence.” IEEE Intelligent Systems, 2010, Vol. 25(6), pp. 13-16.

[5] “The sprout social index, edition IX: Holiday season,” Sprout Social, 02-Nov-2020. [Online]. Available: https://sproutsocial.com/insights/data/q4-2016/. [Accessed: 18-Dec-2021].

[6] G. Silahtaroğlu, and N. Alayoglu, “Using or not using business intelligence and big data for strategic management: an empirical study based on interviews with executives in various sectors.” Procedia-Social and Behavioral Sciences. 2016, Vol. 235, pp. 208-215.

[7] M. Anandarajan, A. Anandarajan, and C. A. Srinivasan, “Business intelligence techniques: a perspective from accounting and finance.” Springer Science & Business Media. 2012, pp. 34-41.

[8] H. Liu and P. Maes, “InterestMap: Harvesting Social Network Profiles for Recommendations.” Proceedings of workshop: Beyond Personalization San Diego. 2005, Pp. 54-59.

[9] 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. 2003, pp. 105-112.

[10] 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. 2003, pp. 129-136.

[11] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, “The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation.” ACM Transactions on Management Information Systems (TMIS). 2018, Vol. 9(2), pp. 1-29.

[12] O. Adwan, M. Al-Tawil, A. Huneiti, R. Shahin, A.A. Zayed, and R. Al-Dibsi, “Twitter sentiment analysis approaches: A survey.” International Journal of Emerging Technologies in Learning (iJET) 15.15 (2020): 79-93.

[13] C. D’Arconte. “Business Intelligence applied in Small Size for-profit companies.” Procedia computer science. 2018, Vol. 131, pp. 45-57.

[14] B. P. Pokharel, “Twitter sentiment analysis during covid-19 outbreak in Nepal.” Available at SSRN 3624719 (2020).

[15] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision.” CS224N project report, Stanford. 2009, Vol.1(12), p.2009.

[16] “Sentiment140 dataset with 1.6 million tweets.” Kaggle, 14-Sept-2017. [Online] Available: https://www.kaggle.com/kazanova/sentiment140 [Accessed: 15-Jan-2022].

[17] L. Ramadhan. “TF-IDF Simplified- Towards Data Science.” Medium, 28-Dec-2021 [Online]. Available: https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530 [Accessed: 25-Jan-2022].

[18] A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, and R. Socher. “Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization.” 2020, arXiv preprint arXiv:2006.09595.

[19] T. M. Alam, and M. J. Awan, “Domain analysis of information extraction techniques.” Int. J. Multidiscip. Sci. Eng 9 (2018): 1-9.

[20] D. M. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness, and correlation.” arXiv preprint arXiv:2010.16061 (2020).

[21] L. Fan, M. Luo, and J. Yin, “Flag choice and Port State Control inspections-Empirical evidence using a simultaneous model,” Transport Policy, vol. 35, pp. 350–357, 2014.

[22] S. Patel. “Chapter 2: SVM (Support Vector Machine) — Theory - Machine Learning 101.” Medium. 10-Nov-2018 [Online] Available: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 [Accessed: 27-Jan-2022].

[23] A. Field. “Discovering Statistics Using SPSS (Introducing Statistical Methods S.)” (2nd Edition) 2005 [E-book]

Authorized licensed use limited to: University of Arkansas Litte Rock. Downloaded on March 07,2023 at 04:31:01 UTC from IEEE Xplore. Restrictions apply.