ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 11 (2021) No. 3, pp. 27-40
DOI:10.33168/JSMS.2021.0302

An Approach to Enhance Business Intelligence and Operations by Sentimental Analysis

Saai Mahesh Srinivasan[1], Preksha Shah[2], Snehitha Sai Surendra[3]

[1, 3] Department of Operations & SCM, International School of Business and Research, Bangalore, Karnataka, India.

[2] Department of Computer Science, University of Massachusetts Lowell, Massachusetts, United States of America

saimahesh.pg20143@isbr.in[1], preksha_shah@student.uml.edu[2], snehitha.pg20135@isbr.in[3]

Abstract. Sentimental analysis is rapidly getting inducted into businesses as a direct result of the technology growth in every sector owing to globalization and industry 4.0. Sentimental analysis which is also known as opinion mining is used in identifying and analyzing text based on the tone that was conveyed by the person which can be categorized broadly into positive, negative and neutral. Businesses can utilize sentimental analysis to tap insight important insights regarding companies, organizations, people, trends and services. With the vast amount of Big Data increasing every day, especially from social media such as Twitter, Facebook etc. businesses can utilize sentimental analysis. This paper thus focuses on implementing machine learning models in Python to perform sentimental analysis from twitter tweets as a viable approach to enhance business intelligence, improve decision marking and target effective operations. The data used in this analysis is obtained from Kaggle collections of COVID-19 twitter dataset. This paper also discusses the various types of applications for sentimental analysis in business and their benefits. The findings from this paper will help improve understanding sentimental analysis for businesses and their practicality in real world scenarios as Big Data advances whilst business intelligence of companies rigorously demands outshining competitive advantage.

Keywords: Business operations, information systems, business intelligence, artificial intelligence, sentimental analysis, natural language processing.

1. Introduction

Sentiment analysis is known has the process of determining whether a text or piece of data is either positive, neutral or negative. Sentiment analysis systems are widely being introduced into businesses due to their high accuracy reliability into combining the techniques of machine learning and natural language processing to allot weighted sentiment values to topics, trends, categories etc. to a text. It is a type of data mining also known as opinion mining wherein text analysis is performed to extract and thus analyze valuable information for business intelligence, which mostly originates from social media and other sources such as news, blogs and so on. The analyzed data would categorize the public opinion or sentiment towards certain happenings in a specific part of the world or even globally that could elaborate on the contextual exploration of the information derived. Sentimental analysis would help data analysts within businesses to capture the open and widely available public data and conduct metrics of market research, business environment scanning, brand image opinion etc. Business is growing increasingly as big data grows for data analytics companies who integrate their own sentiment analysis systems using APIs for social media monitoring, customer satisfaction, employee analysis and customer sentiment analysis. Thus, business intelligence would grow exponentially with sentimental analysis businesses would be able to track the brand perception, product perception, company (brand) reputation which would also create effective methods to improve and expand on the operations of the company. Though sentiment analysis is broadly classified into machine learning and lexicon based, we focus on machine learning using python libraries in this paper. Not only is python programming widely used across the world for data science and business analytics, but with its ease of access in implementing machine learning models and creating important visualizations of data makes it much more convenient and simpler for handing massive amounts of big data. The main role of machine learning for its use in sentimental analysis is to advance and create automation for text analytics. Through the use of machine learning libraries widely available in python such as Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Random Forest classifier, Extreme Gradient Boosting (XGBoost), Logistic Regression and Catboost Model we derive the accuracies and try to understand which algorithm would be best suited as an effective model for sentimental analysis of our dataset. The dataset in use for this paper is the COVID-19 tweets from twitter that is available as a dataset on the website ‘Kaggle’. COVID-19 has plagued the world since 2020 with the outbreak of the pandemic still existing as of today. The feasibility of choosing this dataset serves an important insight on how businesses can make use of such a situation where supply chain, retail and logistics management, businesses and the global market have been crippled in all aspects. Customers are more scattered and divided due to virtualization of almost all modes of shopping and businesses driven to loss with major corporations trying to dominate the e-market space calls for techniques such as sentimental analysis for smart business intelligence. COVID-19 tweets from twitter are vast amounts of data with great potential to analyze and extract information that give companies a competitive advantage for this business. In order to show the importance of advances in Industry 4.0 as well, Twitter serves a good source to obtain unstructured data from the tweets. Due to the vast number of sentiments and various data attributes available from this dataset, sentimental analysis would make it efficient to understand insights into boosting business intelligence and operations. This paper also explores into the applications and advantages of implementing sentimental analysis for business intelligence and operations, which would explore a vague idea of its practicality in implementing such systems for their business.

2. Literature Review

The sentimental analysis is one of the modern tools used to analyze the data from the user's perspective which provides an insight into making decisions for development of businesses. The following studies have become the basis for our study:

"Business intelligence analytics using sentiment analysis-a survey": Prakash P. Rokade and Aruna Kumari D (Feb, 2019) conducted a research on how sentimental analysis used as business intelligence analytics. In this paper the author tells us that the sentimental analysis is fruitful to the individual, business entities and Government to take decisions. This study provides fundamentals and techniques used to extract the sentiments from the text data. This research paper helps us to know the basics of sentimental analysis and it's important for business development.

"Topic based Sentiment Analysis for COVID-19 Tweets": Manal Abdulaziz, Alanoud Alotaibi, Mashail Alsolamy and Abeer Alabbas (2021) conducted a research on tweets during first wave of corona virus. This study is conducted in two different periods where the data is extracted and analyzed the sentiments in tweets posted by people during these two periods. This study become the basis for our study to know the extraction and analyzing the data.

"Global Sentiment Analysis of COVID-19 Tweets over Time": Muvazima Mansoor, Kirthika Gurumurthy, Anantharam R U, and V R Badri Prasad (10th November 2020) conducted a reasearch on how the sentiments of people across different countries had changed with respect to COVID 19 tweets. This study analyzed the sentiments of people on basis of their daily aspects like work from home and online learning that led to the change in sentiments of people over the time. This study gives us an insight of analytical tools that can be used to do sentimental analysis.

"A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis": Furqan Rustam et. al (Feb 25th 2021) conducted a realistic assessment on COVID-19 tweets to know the sentiments of people. This study also analyzed the decisions taken during the pandemic. In this study machine learning approach is used as an analytical tool. The performance of ML models are comparatively studied to understand their accuracies.

"Sentiment Analysis on the Impact of Coronavirus in Social Life Using the BERT Model": Mrityunjay Singh, Amit Kumar Jakhar, Shivam Pandey (15th Feb 2021) conducted a study on twitter data sets where people express their opinions on COVID pandemic. This twitter data is analyzed the mental state of people during the pandemic situation by using BERT model. In this study 2 data sets are used one is from the tweets from other countries and other is from India. This study become the basis of how to do a sentimental analysis for a data set to acquire accurate results.

"Sentiment Analysis of COVID-19 Epidemic Using Machine Learning Algorithms on Twitter": Sudheer Kumar Singh, Dr. Prabhat Verma Dr. Pankaj Kumar (2020) analyzed the tweets to know the sentiments of people during the pandemic. The tweets involve the Challenges faced by people and the sentiments and are predicted as positive, negative and neutral. These sentimental analysis helps us to protect the people from the disease.

"Social media sentiment analysis based on COVID-19": László Nemes and Attila Kiss (14th July, 2020) wrote an article that concludes and analyses the sentiments and manifestation of the users on Twitter platform based on the current trends. This article involves analyzing, compiling, visualizing statistics and summarizing the sentiments and manifestations. This study gives an insight of the tools used to filter the data.

3. Research Methodology

3.1. Dataset Description

The dataset that has been used for this paper was taken from Kaggle's "COVID-19 NLP Text Classification" dataset. These tweets were pulled from Twitter and it has been tagged manually. The dataset has 6 attributes/columns, which are – UserName, ScreenName, Location, TweetAt, OriginalTweet, and Sentiment. The total number of rows of data in the dataset is 41158.

a. UserName – Twitter user name.
b. ScreenName – Screen display name which is usually the real name of user.
c. Location – Location of user.
d. TweetAt – Full date of tweet in DD-MM-YYYY format.
e. OriginalTweet – Tweet written by the user.
f. Sentiment – Sentiments in form of "Extremely Positive", "Positive", "Neutral", "Negative", "Extremely Negative.

3.2. Proposed System Model

The proposed system model for this paper can be divided into various stages that can be better understood through phases.

i. Review Dataset Phase: Reviewing the feasibility of the COVID-19 dataset from Kaggle to check whether it would be compatible to our objectives of finding important insights such as common words in the dataset, common words based on sentiment, hashtags in a tweet from the user, trends association with the tweets, sentimental trends and how they can affect businesses and can be used for market research, competitive analysis and business environment analysis. This phases helps to clear up any null values, understand process of sentiments, identify top locations from the tweets originate and visualize them for our understanding, identifying missing values etc.

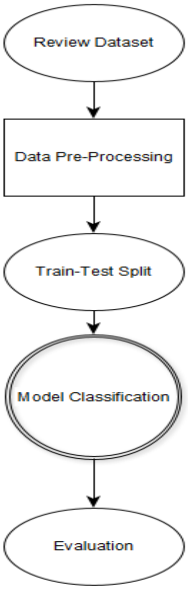

Fig. 1: Proposed System Model.

ii. Data Pre-Processing Phase: For the model to get fed with accurate data, it is require to remove unnecessary inputs and texts so in this phase the data is preprocessed before using it for the model i.e converting raw data into clean data to make it feasible for our analysis. This phases removes @user, http, urls and other special characters. Punctuations, numbers, special characters, stop words are removed. Then, tokenization is done to create tokenized tweets.

iii. Train-Test Split Phase: In this phase the data is divided into subsets of training and testing. The division of these training and testing subsets are in the ratio of 80% and 20% respectively.

iv. Classification Phase: The models such as SGD, Random Forest, XGBoost, Catboost, Logistic Regressio and SVM are implemented for the multiclass classification. It is required to derive a higher accuracy than what the multiclass classification would obtain and thus the models have to be tuned towards the type of data being implemented i.e transform into binary classification in a one-against-all technique.

v. Evaluation Phase: This phase finally gives an evaluation of all the models implemented and also gives us the results for our analysis. The best model through train and test will be obtained in this phase.

4. Results

4.1. Converting Multiclass Classification to Binary Classification

The OAA (one against all) approach to multiclass classification allows to reduce any multiclass problem to a problem solvable using binary classifiers. It's done to achieve maximum accuracy since for the multiclass classification, the accuracy achieved was around 50% and by using binary classification the accuracy obtained is 80%. We implement this using the CatBoost algorithm. Which is comparatively easier to implement as it's a recent and open sourced ML algorithm from Yandex. It's used for categorical features and boost through gradient boosting, which gives it the name CatBoost. After converting it from multiclass to binary through Natural Language Toolkit (nltk) stopwords and nltk corpus, we apply lambda for stop words and get our pre-processed data. As for the CatBoost model, we install the Catboost package and implement the algorithm using CatBoostClassifier function. Just like other models shown below, we train, test split the data and then obtain the training and validation scores.

4.2. Train and Test Data

The train-test split is a method for assessing the performance of a machine learning algorithm. It is applicable to classification and regression issues, as well as any supervised learning approach. The data set is divided into two subsets, the first of which is used to fit the model and is referred to as the training dataset. The model is not trained on the second subset but rather is used for testing the accuracy to that of the training data. Here, the main parameter is the size of the test and train sets for the data. Depending on the type of dataset and data analysis that is to be performed, the common split percentages include:

- Train: 80%, Test: 20%
- Train: 67%, Test: 33%
- Train: 50%, Test: 50%

Python makes it possible through the scikit learn ML library to provide an implementation of the train-test split procedure with the train_test_split() function. To accomplish our objective, we stick to 80% train and 20% test splitting of data.

4.3. Count Vectorizer

Scikit-learn has a function called as CountVectorizer( ) that is used to convert a collection of text documents to a vector of term/token counts. This is possible by enabling the pre-processing of text data prior to generating the vector format of representation. The functionality makes it a very flexible feature for a text representation module

The CountVectorizer function forms a matrix where each unique word is represented by a column of the matrix and each of the text sample from the document is a row in the matrix. The count of the word is the value of the cell for the respective text sample.

4.4. Logistic Regression

Logistic regression is a statistical machine learning technique that classifies data by taking the outcome variables at their extremes and attempting to draw a logarithmic line that separates them. The simplest regression analysis is the binary variable regression analysis. The model is: $Y=\beta 0 + \beta 1Xi + u$ ln this model, while Y represents the dependent variable, X shows the independent variable; β0 the invariable; β1 obliquity; and u the margin of error. We imported the model from sklearn.linear_model for our implementation. We then fit the data from the train and test split.

4.5. Support Vector Machine

Another way for classifying problems such as breast cancer is the support vector machine, or SVM. SVM seeks to separate the data using a 'separating hyperplane', and it performs well on datasets that are linearly separable. The SVM model is a representation of instances as points in space that is mapped so that the examples of the different categories are separated by a distinct gap that is made as large as possible. These new examples are then mapped into that space, which aids in predicting which category they based on which piece of the gap they fall into. The methodology for SVM was to setup the data frame and then fit the data with the train and test split.

4.6. Random Forest

Random forest is a meta-based estimator that employs averaging to increase forecast accuracy and control over-fitting by combining many decision tree classifiers on different sub-samples of the dataset. We use sklearn, ensemble. RandomForestClassifier from scikit-learn package to do the classification. In order to build a forest of trees from the training set (X,y) and predict(X) to predict class for X the function fit(X,y) was used. Another method that could be used for stratifying the dataset was stratified dataset. When a dataset is very highly skewed with a particular label field. It can affect the performance of that model because the dataset the model was trained on is not reflective of the dataset the model was tested on. When a dataset is stratified, the dataset is split in a way where the class distribution of the train dataset is roughly equal to the dataset. We then fit the data from the train and test split.

4.7. Stochastic Gradient Descent

Instead of using the whole gradient with all available data, Stochastic Gradient Descent (SGD) is an iterative based optimization technique that employs tiny groups/bins of data to construct an expectation of the gradient. We have the following for weights w and a loss function L: wt+1=wt−η∇^wL(wt), where η is a learning rate. When opposed to batch gradient descent, which involves recompilation on gradients for similar examples before each parameter update, SGD helps to reduce redundancy, making it significantly faster. We use sklearn.linear_model to import SGD regression model using import SGDClassifier. We then fit the data from the train and test split.

4.8. Extreme Gradient Boosting

Extreme gradient boosting which is also known as XGBoost is built on the principles of gradient boosting framework. It is designed to push the computation limits of computation to extreme limits in order to provide a scalable and accurate library. Gradient boosting is a ML technique for both regression and classification problems. It produces a prediction model in the form of an ensemble of decision trees. Here, the model is built similar to a stage-level fashion, which is is done by other boosting techniques. XGBoost performs an optimization of differentiable loss function. XGBoost also helps reduce over fitting. We then fit the data from the train and test split.

4.9. Evaluation of Model Accuracies

The highest accuracy achieved was by using Logistic Regression Model and it is 80.6%. The model we decided on using in order to test accuracies are Logistic Regression, Naïve Bayes, CatBoost, SGD, Random Forest, XGBoost, SVM. We then summarize all the accuracies by using pandas dataframe to combine the variables set for test accuracy and use a user defined text of ‘Model ’as a column to classify it respective to the model. The accuracies in order of descending order are shown in the table below:

Table 1. Model Accuracies Summarized

Model	Accuracy
Logistic Regression	80.6%
CatBoost	79.6%
Stochastic Gradient Descent	79.2%
Random Forest	78.7%
XGBoost	78.6%
Support Vector Machines	77%

4.10. WordCloud Generation

WordCloud is a collection or cluster of words, and it can be generated in python by using WordCloud( ) importing from WordCloud.

We used the following keywords “operations, logistics, supplychain, warehouse, manufacturing, distribution, e-commerce, shipping, delivery, retail, transportation, inventory, supplier, procurement” to generate wordclouds around the 5 categories of reactions on tweets i.e., extremely positive, positive, extremely negative, negative, and neutral.

Fig. 8: WordCloud for Extremely Positive Reaction.

Fig. 9: WordCloud for Positive Reaction.

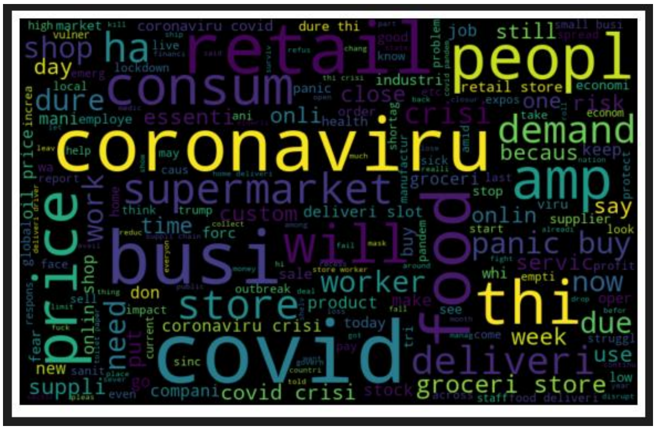

Fig. 10: WordCloud for Extremely Negative Reaction.

Fig. 11: WordCloud for Negative Reaction.

Fig. 12: WordCloud for Neutral Reaction.

5. Discussion

Practicality of implementing achieved results in real world business scenarios through the use applications of sentimental analysis in business can't be accessed due its immense value of various applications in business scenarios. The complete brand revitalization can be demonstrated through sentiment analysis in business. The success of business with the sentiments data is the potential use of unstructured data for actionable conclusions. Machine learning models are manually created features before classification that served as the purpose for past few years. Thus, AI has significantly helped in the following for sentimental analysis as well:

- Automatically extracts applicable features.
- Helps to drag off the redundant features.
- Eliminates additional efforts of manually crafting the features.

5.1. Business Intelligence

The elimination of guess work and accomplishing the timely decisions can be done with the help of understandings that drawn from rich information. The customer retention rate on newly established product can be easily estimated through sentiment data. The reviews generated from sentiment analysis in business is helpful to survive in present market and satisfy customers. If the results are automated and it helps management to take immediate decisions. Staying dynamic in this competitive world throughout their journey is considered as business intelligence. And it is achieved by having the sentiments data. The new idea can be verified before bringing it into reality. This is referred as concept testing. The new product, campaign or a new logo can be analyzed by keeping it into concept testing where it analyzes the sentiments attached to it.

5.2. Competitive Advantage

The elimination of guess work and accomplishing the timely decisions can be achieved with the help of sentiments data. One should be open to experiment with it tactfully to gain on the business applications of sentiment analysis. Like it is already known that a piece of text is used to perform sentiment analysis. Applying just to own brand makes it success? X% that deals with reviews (positive or negative) and y% metric to compare them with helps to survive in competitive markets. To gain opportunity and to improve performance of business one should know about their competitor's data. Current customer trends can be predicted with the help of sentiment analysis. New strategies can be easily developed by acquiring the current trends to compete in the market on newly established and products, and also helps to estimate customer retention rate. In order to adopt the current market situation and satisfy customer's reviews of sentiment analysis can be helpful in better way. Automated conclusions can lead to take immediate decisions. With the help of sentiment data one can own business intelligence to stay dynamic in the market. A new product, campaign and logo can be tested by using concept testing to analyze sentiments attached to it.

5.3. Customer Experience

A business can be succeeded once it reaches to the maximum satisfaction of customers. The customer experience can be either positive, negative or neutral. During this internet era, the experience of customers can be expressed by posting socially and feedback online. The sentiments attached to this data can be detected and categorized according to the tone and temperament. This detection of data helps to know proper implementation and improvement of products, services and customer support can be easily made. Positive response from customers in not enough to sustain but a proper customer system of company should be available to customers regardless of the appeal to the customers that the systems provides, this helps to sustain in market for long-term.

5.4. Brand Reputation

The products that manufactured and the services it provides doesn't defines brand of a company. The name and fame of a brand can be built with the help of social media campaigning, online advertising, content writing, digital-marketing and customer support systems. The perception of present and potential customers can be quantified with the help of sentiment analysis in business. The negative sentiments foster a better appeal of brand techniques and other strategies for market analysis to improve brand status to a more prominent level. A quick transition can be seen in business with the help of sentiment analysis. The business applications of sentiment analysis are many and immense. A greater business value can be achieved depends on the tool used and how well it is used can be more advantageous to sustain.

6. Conclusion

In conclusion, in view of the future work and potential development of this research, it can be advised to create a visual or graphical interface that helps the user of the system interact in a much more informative manner. Through the help of various databases, ERP system integrations, blockchain technologies, classifications of various data types based on the business etc. can be considered. Besides sentimental analysis, it is possible to add on better tensor flow features and ensuring that the packages and tools used are up-to-date. Based on the business intelligence required, the complexity of a similarly proposed system in this paper can be implemented.

However, in this paper we've identified that the best model for our data analysis is 'Logistic Regression' with an 80% accuracy rate. This could mainly be attributed to the logistic regression implementing its function on deriving better solutions to classification problems wherein the dependent categorical variable is predicted from the independent categorical variable. Twitter has shown that it has huge potential as a source of Big Data for companies to perform such types of emotional or sentimental analysis.

Our results have obtained an important overview of how successfully a business or organization can capitalize on upgrading their technology prowess in the automation or nearly automated systems with integration of artificial intelligence through machine and deep learning techniques. This compliments that Big Data from Twitter's tweets can be used for global or local environments for business intelligence activities. These results have also helped us understand the various practical approaches and applications of sentimental analysis for businesses. Thus, it is also important to look forward to analyzing social media as a source of Big Data, where companies such as Facebook and the companies it owns like Instagram, WhatsApp etc. would prove to be very beneficial as there will be a wide variety and diversity of data to work on for insights.

The limitations to this paper is that the disadvantages and possible implications of their implementation such as manpower, cost, technological complexity etc. aren't discussed. BERT model which is known to have a better accuracy hasn't been implemented due to cost and time constraints. However, this could be taken as a research gap for anyone looking to continue research in this domain.

References

Applications of Sentiment Analysis in Business. (2018). https://towardsdatascience.com/applications-of-sentiment-analysis-in-business-b7e660e3de69. Accessed 9th June, 2021.

COVID-19 Dataset. (2020). https://www.kaggle.com/datatattle/covid-19-nlp-text-classification. *Accessed 1st June, 2021*.

Quang H. N. & Huu T. D. (2019). Factors Preventing the Way to Success of the Retail Supply Chain. *Journal of System and Management Sciences*, 9(2), 114-122.

Kim J. B. (2019). Implementation of Artificial Intelligence System and Traditional System: A Comparative Study. *Journal of System and Management Sciences*, 9(3), 135-146.

Mahmood A. T., Kamaruddin S.S., Naser R. K., Nadzi M.M. (2020). A Combination of Lexicon and Machine Learning Approaches for Sentiment Analysis on Facebook. *Journal of System and Management Sciences*, 10(3), 140-150.

Manal A., Alanoud A., Mashail A. and Abeer A. (2021). Topic based Sentiment Analysis for COVID-19 Tweets. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(1).

Mansoor, M., Kirthika G., and V. R. Prasad. (2020). Global Sentiment Analysis of COVID-19 Tweets over Time. *arXiv preprint arXiv: 2010.14234.*

Nazifa T. H. & Ramachandran K.K. (2018). Exploring the Role of Information Sharing in Supply Chain Management: A Case Study. *Journal of System and Management Sciences*, 8(4), 13-37.

Nemes L. & Kiss A. (2020). Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*, 5(1), 1-15.

Prakash R., & Aruna D., (2019). Business intelligence analytics using sentiment analysis-a survey. *International Journal of Electrical and Computer Engineering (IJECE)*, 9(1), 613-620.

Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi G.S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS ONE*, 16(2*).*

Sentimental Analysis Explained. (2018). https://www.lexalytics.com/technology/sentiment-analysis. Accessed 10th June, 2021.

Singh, M., Jakhar, A.K. & Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Soc. Netw. Anal. Min*. 11(33).

Singh S. K., Verma P., Kumar P. (2020). Sentiment Analysis of COVID-19 Epidemic Using Machine Learning Algorithms on Twitter. *Journal of Critical Reviews.* 7(18), 2565-2572.