Research Article

Sentiment Analysis Of Twitter Data By Using Deep Learning And Machine Learning

¹Prof. Manisha Sachin Dabade, ²Prof. Malan Dipak Sale, ³Prof. Dhanashri D. Dhokate, ⁴Prof. Shweta M.Kambare

¹Assistant Professor, Department of CSE, Sharad Institute of Technology College of Engineering, Ichalkaranji (Kolhapur)

²Assistant Professor, Department of CSE, Sinhgad College of Engineering, Pune.

³Assistant Professor, Department of CSE, Padmabhooshan Vasantraodada Patil Institute of Technology, Budhgoan.

⁴Assistant Professor, Department of CSE, Sinhgad College of Engineering, Pune

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract — In today's world, social media is viral and easily accessible. The Social media sites like Twitter, Facebook, Tumblr, primary and valuable of information. etc. are а source Twitter is a micro-blogging platform, and it provides an enormous amount of data. Such type of information can use for different sentiment analysis applications such as reviews, predictions, elections, marketing, etc. It is one of the most popular sites where peoples write tweets, retweets, and interact daily. Monitoring and analyzing these tweets give valuable feedback to users. Due to this data's large size, sentiment analysis is using to analyze this data without going through millions of tweets manually. Any user writes their reviews about different products, topics, or events on Twitter, called tweets and retweets. People also use emojis such as happy, sad, and neutral in expressing their emotions, so these sites contain expansive volumes of unprocessed data called raw data. The main goal of this research is to recognize the algorithms by using Machine Learning Classifiers. The study intends to categorize Fine-grain sentiments within Tweets of Vaccination (89974 tweets) through machine learning and a deep learning approach. The study takes consideration of both labeled and unlabeled data. It also detects emojis from tweets using machine learning libraries like Textblob, Vadar, Fast text, Flair, Genism, spaCy, and NLTK.

Keywords— Python, machine learning, deep learning, fine grain sentiment analysis

I. INTRODUCTION

Sentiment analysis is an automation process. It divides writing pieces into positive, negative, or neutral types. It has an information extraction process that gives answers to questions on public opinions and summarizes several people's viewpoints [9]. Such an approach uses various fields like sociology, psychology, political science, making policies, business analytics, law, etc. The social media posts might never post in a language comprehensible to everyone-English [8]. It is essential to expand techniques and tools to develop sentiment analysis that covers languages that are not well known.

Sentimate Analysis aims to tackle this data to gain important information about public opinion. This information would help to make smarter business decisions, political campaigns, and better product consumption. The textual information retrieval techniques mainly focus on searching, processing, or analyzing the actual data present. It has different applications that help in marketing, business and increases sales of the product people. In the sentence analysis, all the sentences are taken into the .csv file, preprocessed those sentences, and using a machine learning algorithm, the data classified.

1.1 Aim and Objectives

The research's primary goal is to evaluate and improve the sentiment analysis model's performance using machine learning libraries such as TextBlob, VADER, FastText, Flair, Gensim, spaCy, NLTK on a Twitter dataset of Vaccination.

Objectives:

- To compare the metrics obtained from different machine learning algorithms, based on the dataset size.
- To identify the best algorithm suited for sentiment analysis.

1.2 Research Questions

The research methodology needs to answer the following questions:

1. Which machine-learning algorithm uses for analysis and performance evaluation of Fine-grain sentiment analysis on Twitter data?

2. Which is the best fit classification algorithm for conducting the Sentiment Analysis on Twitter data?

3.How the performance of the machine learning algorithms varies as the quantity of labeled and unlabeled data changes?

1.3 Ethical Considerations

The recording details used in this study will remain private and never reveals to a third party. All the participants will not compel to share their personal information, and the copyright standards consider.

II BACKGROUND AND LITERATURE REVIEW

The research gives levels of sentiment analysis. The paper also gives several approaches to sentiment analysis. The different techniques and methods used to implement sentiment analysis are explained below:

A. DATA MINING

The evolution of information technology leads to the growth of data management. The data has grown beyond the levels of databases, leading to an increase in data warehouses. Data mining is the extraction of knowledge from a large amount of data. The knowledge brings out by come across the data patterns. The data mining tools are used to predict knowledge-focused decisions. This becomes a trend in the future and behaviors of businesses.

The decisions driven are proactive and knowledge-driven. Data mining provides retrospective tools of decision support systems to offer prospective analysis automation and move beyond the study of past events. The data mining tools assist to answer business-related questions which are time-consuming to sort out. So most of the companies collect and refine enormous quantities of data forehand [16].

B. TEXT MINING

Text mining is different from data mining. Text mining discovers unknown information while data mining does not. It extracts the patterns from the natural language text [19]

C. SENTIMENT ANALYSIS

Sentiment Analysis is a subfield of text mining, and it's also called opinion mining. It's a way to predict people's feelings or emotions towards a subject or an entity. The analysis can be done in various forms such as Natural Language Processing methods, applying lexicons with annotated word polarities with machine-learning approaches. [13]. The proliferation of user-generated content on the web has resulted in active research on sentiment analysis [13][14]. The performance of sentiment analysis has published by Ghiassi and S. Lee [2].

The sentiment analysis is a classified problem in which texts should be marked as neutral, positive, or negative [2]. There may be regression problems where continuous output is required. In any case, the format of the primary challenges and applications inner of the estimation investigation field was set out by Jianqiang, Gui, and Zhang X [4]. The researcher has provided subjective analysis and researched sentiment analysis as a tool. It recognizes, extracts, and interprets opinions irrespective of various topics within a provided textual context [4].

The authors have projected a clustered algorithm. It classifies adjectives into multiple orientations—much more research done on sentiment analysis. A study [7] revealed that The log-linear model categorizes semantic orientations of conjoined adjectives. On the other hand, the author has employed a web search engine. The point-wise mutual information about the categorization of words used. These included nouns, verbs, adjectives, and adverbs and utilizes predefined linguistic heuristics and seed words. The past works of [9] concentrated on sentiment classification of the whole document that has involved two distinct models.

First, the model utilizes cognitive-linguistic while others utilize construction of manual or lexicon words [6]. The authors attempted to identify characteristics used to recognize whether a given document possesses an emotional language or not[4]. The authors have worked with Machine learning algorithms. Some of the Machine Learning algorithms are Naive Bayes, Random Forest, and XG Boost, CNN-LSTM, and CNN-LSTM. The model classifies the sentimental analysis of the Twitter database. It has the highest accuracy [11].

D. FINE GRAIN SENTIMENT ANALYSIS

Fine-grained sentiment analysis gives correct results to the open conclusion concerning the subject and its more challenging task. It uses five discrete classes, Very Negative, Negative, Impartial, Positive, and Very Positive. Fine-grained sentiment analysis refers to the detection of sentiment. It is not only on the document or post level but also on the sentence, sub sentence. This allows for more finer-grained output.

It is useful for recognizing sentiment that may not be collected by document-based sentiment analysis models. It has linguistic rules, or aspect-based sentiment analysis, related to the sentiment or the feature of a product. The fine grain sentiment analysis can be done using rule-based methods or feature-based methods or by using embedded-based methods.

E. APPROACHES TO MACHINE LEARNING

Machine learning is a technique is used to turn given information into knowledge. This numerous data is useless unless it analyses data and finds the patterns hidden within this data. Machine learning approaches or techniques are used to find different valuable patterns within huge complex data. The hidden patterns and knowledge about a problem are helpful for prediction. It performs all kinds of complex decision-making and techniques broadly employ in sentiment analysis. Alharbi and Elise used the movie reviews dataset to categorize the general sentiment reviews [1]. It concludes that machine learning techniques generate excellent relative performance in support of sentiment task classification. These matched the baselines caused by humans. Their study revealed that the best performance of various classifiers of machine learning gave by the support vector machine (SVM), and the worst performance given by the Naïve Bayes classifier.

Giachanou and Fabio tried machine learning techniques to use machine learning techniques for sentiment analysis. It takes domain and time by considering matching data. Giachanou and Fabio[3] both utilized emoticons to

create labeled data for training. This is to ensure that training data is independent of time, domain, and topic. The research concludes that the subject, style of language, and field are dependent on the sentiment classification task. The researchers have used a machine learning approach with a variation model and feature extraction in support of the text classification task [5].

The authors concluded that the use of bigram features leads to excellent results in sentiment analysis. The sentimental study for short snippet tasks establishes by authors such as Wehrmann, Willian, and Rodrigo using Naïve Bayes. These generated better results as matched to support vector machine [10]. When the document length has increased, the achieved result turned the opposite. The researchers have opted to utilize various sentiment analysis features. E.g., Part of Speech (POS) tagging, a high order of n-grams, etc. In the process of syntactic analysis, the POS tagging has given a demand, leading to an accuracy of about 90% [10].

F. SENTIMENT ANALYSIS ON TWITTER

Social media sites like Twitter, Facebook have various posts for sentiment analysis. Sentiment analysis is a significant point of interest for many researchers. Severyn and Alessandro focused on solving queries like steered classification tasks [8].

Machine learning techniques are hanging on the supervised classification methods. Sentiment recognition is a binary which is used values as positive, negative, and neutral [14]. In this approach labeled data is required for training classifiers [16]. In this approach, the feature of a word local context is required to be taken into accounts like value is negative (e.g., not beautiful) and intensification (e.g., Very beautiful) [15]. The authors attained an accuracy of 80 % using emoticons in labeling tweets and this lead to a noisy corpus. The corpus contained 300,000 tweets which are labeled as positive, negative, and neutral.

The researchers have used newswire tweets data that have happy and sad emoticons queries for creating that corpus and used POS tags coupled with sentiments tweet labels. In the paper [7], Severyn, A., and Alessandro used speech features, which do not improve classifying Twitter messages for sentiment analysis. Severyn and Alessandro [7] created a sentiment analysis on Twitter using spaCy and NLTK tools. These tools were used for the analysis of tweets that are associated with the presidential election in the USA.

G. DEEP LEARNING

Deep Learning is a Machine Learning methods class. Deep Learning is based on artificial intelligence. It extracts high features progressively by using multiple layers. The arrival of deep learning techniques has opened

new doors for fresh possibilities and horizons. Deep understanding requires deep neural networks in learning multifaceted features. These features have been extracted from the assistance of minimum external contributions [7]. The deep learning approach required a large dataset to get a significant boost in the performance of a model. Le, BAC, and Nguyen utilized deep learning techniques to classify sentiments in the review of movies [5].

The researchers have utilized two different datasets that have binary and multiclass labels. The methods of deep learning used recursive neural networks and word2vec both. Such tools were used to create feature vectors in support of classifiers. The aim of the study is at extracting semantic characteristics ad of the texts that matched other previous studies. Word2vec library is used for creating vector representations in words in higher dimensions. So, the authors eased the extraction of deep semantic associations in terms. The authors have used datasets that entailed comments on Chinese Clothing products and attained an accuracy of 90%. Similarly, Ramadhan and Hong [6], employed a model of an external neural network. They used word2vec for the classification of sentimental posts. The posts are from Chinese Sina Weibo.

III. METHODOLOGY

A. DATA COLLECTION

The first phase of machine learning or deep learning implementation comes as data collection. The dataset for the implementation of the sentimental analysis is easily available on the Kaggle. The data collection technique is the sampling of a dataset. The sample dataset is downloaded from the website i.e., the Kaggle website. The dataset used for the implementation of this project is the twitter-vaccination-dataset. The dataset has tweets and retweets which are in raw form. Almost 89900 tweets are collected for implementation.

B. CLEANING OF DATASET

Libraries like pandas, NumPy are available for data preprocessing. Data cleaning or preprocessing is a step-bystep process. The cleaning of the dataset includes various operations like Data cleaning, Data Integration, Data Transformation, and Data Reduction. It includes identifying missing data, ignoring the tuples, filling the missing tuples, removing noisy data that is meaningless. The next step comes as data transformation which includes Data Reduction, Attribute Selection, Hierarchy Generation, Discretization, Normalization, Dimensionality Reduction, etc.

C. IMPLEMENTATION

The previous model has built and trained using the CNN-LSTM algorithm [11]. It gave a training accuracy of 96.5 % and a testing accuracy of 88.1% before.

By using Fine-grain sentiment analysis, the study will focus on the improvement of the accuracy of the model. The model improvement will use the following python programming language and libraries needed for the improvement of the model:

• Python

Python is an interpreter. Python is a high-level as well as an object-oriented scripting language. It is highly readable. Python has fewer syntactical constructions as compare to other programming languages. Python programming is used in the development of the model. In this research work, the following python libraries will be used to develop the machine learning models:

• Pandas

Pandas is a Python package. It acts as a data analysis tool also deals with data structures. Pandas are used for data analysis workflow. It does not need to switch to a more domain-specific language such as R programming language.

• Numpy

NumPy is a package in Python. Numpy having a large collection of high-level mathematical functions. NumPy library supports multi-dimensional arrays as well as matrices.

• Scikit-learn

Scikit-learn is a machine learning library. This library is available in Python. It is used for data analysis and data mining. By using this library dataset is split into two different datasets i.e., a training dataset and a testing dataset. The accuracy score of the model is also determined using this library.

• Matplotlib

It is a python library that generates plots, histograms, power spectra, bar charts, etc. The graphs are generated using matplotlib.pyplot module. It's a visualization library in python. This library is used for 2Dplots of the array. It is a library of multiplatform data visualization. It is built on NumPy arrays. It is designed for work with the broader SciPy stack.

• TextBlob

It is a popular Python library. This library is for processing textual data. It is built on top of NLTK. It uses a sentiment lexicon that has predefined words. It gives scores for each word. After that, they averaged out using a weighted average to give an overall sentence sentiment score. The tasks may be part-of-speech tagging, sentiment analysis, noun phrase extraction, translation, classification, etc.

• VADER

VADER stands for Valence Aware Dictionary and Sentimate Reasoner. It deals mostly with the different texts of social media, NY Times editorials, movie reviews, and different product reviews. It also uses a mixture of a sentiment lexicon. A sentiment lexicon is simply a list of lexical features. All lexicon ratings are calculated using the Compound score. These ratings have been regulated among

-1(most extreme negative) and +1 (most extreme positive).

Positive sentiment : (compound score>=0.05) neutral sentiment : compound score is in between -0.05 and 0.05 negative sentiment : (compound score <= -0.05)

• FastText

FastText is a mainly CPU-based library used for text representation as well as for classification. FastText think about different subwords with the help of a collection of n-grams: for example, "train" is split into "tra", "rai", and "ain". In this way, the representation of a word is more resistant to misspellings. It is also more resistant to minor spelling variations.

• Flair

Flair is a contextualized representation. It is a PyTorch-based framework. For contextualized representation, sentences from large strings are broken down into different character sequences. These character sequences are pre-trained in a bidirectional language model. This "learns" embeddings at the character-level. Like this, the model learns to disambiguate case-sensitive characters e.g., proper nouns form similar-sounding common nouns and other syntactic patterns into natural language. This pattern makes it very powerful for tasks like named entity recognition and parts-of-speech tagging.

• Gensim

Gensim is an open-source library. It's used for unsupervised learning. It is for subject modeling and natural language processing. Gensim is planned to handle huge content collections utilizing data streaming and incremental online algorithms, which separates it from most other machine learning program bundles that target as it were in-memory processing.

Gensim includes streamed parallelized implementations of word2vec algorithms where,

1. A document is a list of strings

2. A token is an occurrence of a grouping of characters. For the most part, they represent the words and terms in a text.

3. Word embedding could be a multidimensional representation of the content. It is fundamentally changing over the data in the content frame into a number.

4. A dictionary is a mapping of words with an id for each word.

• spaCy

It is an open-source library and provides advanced Natural Language Processing. It's built on the very latest research and was designed from day one to be utilized in real products. It's becoming increasingly popular for processing and analyzing data in Natural Language Processing.

• NLTK

It is a Natural Language Toolkit. It converts text into numbers. This model then easily works with. NLTK works with human language data and the leading platform for building Python programs. This is the powerful NLP library. It has packages that make machines understand human language. It answers it with an appropriate response. NLTK includes Stemming, tokenization, Lemmatization, Character count, Punctuation, word count packages.

IV . IMPLEMENTATION DETAILS

The libraries such as TextBlob, VADER, and the algorithms such as logistic regression, Support Vector Machine(SVM), and the embedding methods such as FastText, Flair are used to enhance the performance of the model. The workflow is as below:



Fig. 1. Workflow Diagram

The algorithms for the implementation of the model are:

A] Logistic Regression and

B] Support Vector Machine [SVM]

A. Logistic Regression:

Logistic regression is a statistical model which utilizes a logistic function to sample binary dependent variables. It has a dependent variable that possesses two feasible values. These feasible values are pass or fail. These values are shown by an indicator variable. The indicator variable has two values named "0" or "1".



Fig.2. Logistic regression model

B. Support Vector Machine(SVM)

It is a type of supervised machine learning algorithm. It is used for classification and regression challenges. But it's mainly used for classification problems. Every data item is organized as a point in SVM. It has ndimensional space and contains the value for each feature. Each feature is of a particular coordinate.



Fig. 3. Support Vector Machine

Simply the coordinates of individual observation are the Support Vectors. The SVM is available in the sci-kit-learn library.

V . PERFORMANCE EVALUATION

The performance of a model is based on accuracy, precision, and recall. These parameters are calculated as below:

Table I:	Confusion	Matrix
----------	-----------	--------

	Predicted Positive	Predicted Negatives
Actual	TP1	FN1
Positive		
Actual Negative	FP1	TN1

Accuracy = (TP1+TN1)/(TP1+TN1+FP1+FN1) Precision=TP1/ (TP1+FP1) Recall=TP1/ (TP1+FN1) F1= (2* Precision*Recall) / (Precision+ Recall)

Where TP true positive instances, FN is false-positive instances and TN is the false-negative instances.

VI. CONCLUSION

In this research, the main focus is on the Fine-grain sentiment analysis of tweets and to study emoji labels. The research will work on both labeled and unlabeled data. The machine learning libraries such as TextBlob, VADER, FastText, Flair, Gensim are used to improve the. All the algorithms are subjected to experiment and then the results are drawn with the performance metrics chosen to address them.

VII. FUTURE WORK

The chosen deep learning and machine learning algorithms are implemented on a specified size of data. The results obtained are subjective to the used dataset. Thus, presenting the scope for future research of validating the above-stated models with different sizes of the dataset and estimating the results. This can also give a hint for future research on analyzing the impact of the size of the dataset on the accomplishment of the Machine Learning algorithms. The above study predicted the text data more accurately when compared to emoji data detection. In this research, labeled data is positive, negative, and neutral type. In future work, consideration of emoji labels and collection of a better and large number of emoji datasets having proper labels may help to improve the results.

REFERENCES

1. Alharbi, A. M., and Elise D. "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information." Cognitive Systems Research 54 (2019): 50-61.

- 2. Ghiassi, M., and S. Lee. "A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach." Expert Systems with Applications 106 (2018): 197-216.
- 3. Giachanou, A., and Fabio C. "Like it or not: A survey of Twitter sentiment analysis methods. " ACM Computing Surveys (CSUR) 49, no. 2 (2016): 1-41.
- 4. Jianqiang, Z., GUI X., and Zhang X. "Deep convolution neural networks for Twitter sentiment analysis." IEEE Access 6 (2018): 23253-23260.
- 5. Le, BAC, and Huy Nguyen. "Twitter sentiment analysis using machine learning techniques." In Advanced Computational Methods for Knowledge Engineering, pp. 279-289. Springer, Cham, 2015.
- 6. Ramadhan, A. M., and Hong S. G. "Twitter sentiment analysis using deep learning methods." In 2017 7th International annual engineering (inaes), pp. 1-4. IEEE, 2017.
- Severyn, A. And Alessandro M. "Unitn: Training deep convolutional neural network for Twitter sentiment classification." In Proceeding of the 9th International Workshop on Semantic Evaluation (semeval 2015), pp. 464-469. 2015.
- Severyn, A., and Alessandro M. "Twitter sentiment analysis with deep convolutional neural networks." In Proceeding of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 959-962. 2015.
- 9. Symeonidis, Symeon, Dimitrios Effrosynidis, and Avi Arampatzis. "A comparative evaluation of preprocessing techniques and their interactions for Twitter sentiment analysis."Expert System with Applications110 (2018): 298-310.
- 10. Wehrmann, J., Willian B., Henry EL Cagnini, and Rodrigo C. Barros. "A character-based convolutional neural network for language-agnostic Twitter sentiment analysis." In 2017 Internatio Convolutional Neural Networks. Master's Theses in Mathematical Sciences.
- 11. Kundan Reddy Manda., "Sentiment Analysis of Twitter Data Using Machine Learning and Deep Learning Methods" Master of Science in Computer Science June 2019, Karlskrona, Sweden
- A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for imdb review sentiment analysis," in 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017, pp. 540–546.
- 13. A Survey of Opinion Mining and Sentiment Analysis | springerlink." [Online]. Available: <u>https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_13</u>.
- 14. T. Carpenter, and T. Way, "Tracking Sentiment Analysis through Twitter,". ACM computer survey. Villanova: villanovauniversity, 2010.
- 15. M.Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon- Based Methods for Sentiment Analysis," Association for Computational Linguistics, 2011.
- P. Goncalves, F. Benevenuto, M. Araujo, and M. Cha, "Comparing and Combining Sentiment Analysis Methods", 2013.
- 17. A.Sharma, and S. Dey, "Performance Investigation of Feature Selection
- 18. Methods and Sentiment Lexicons for Sentiment Analysis," Association for the advancement of Artificial Intelligence, 2012.
- Bhagyashri Wagh, Prof. J. V. Shinde, Prof. P. A. Kale. "A Twitter Sentiment Analysis Using NLTK and Machine Learning Techniques" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-12) 2017
- 20. "Differentiating data- and text-mining terminology." [Online]. Available: https://dl.acm.org/citation.cfm?Id=954024.
- 21. SRIVASTAVA, SRISHTI, et al. "ANALYSIS AND COMPARISON OF LOAN SANCTION PREDICTION MODEL USING PYTHON." International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) 8.2, Jun 2018, 1-8
- 22. MARTÍNEZ, DANIEL, OCTAVIO JOSÉ SALCEDO PARRA, and MARCO ANTONIO AGUILERA. "MACHINE LEARNING APPLIED TO ICFES TESTS TO IDENTIFY MEASUREMENT PATTERNS IN STUDENTS." International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) ISSN (P): 2249–6890; ISSN (E): 2249–8001 Vol. 10.6, Dec 2020, 1–8
- 23. Satyanarayana, K. N. V., et al. "Mobile app & iot based smart weather station." International Journal of Electronics, Communication and Instrumentation Engineering Research and Development (IJECIERD) 7.4 (2017): 1-8.
- 24. Bepery, Chinmay, et al. "Advanced home automation system using Raspberry-Pi and Arduino." International journal of computer science and engineering 8.2 (2019): 1-10.
- 25. MEENAKSHI, S., M. PREMA, and P. VIJAYAPRIYA. "IS THE ACQUISITION OF ENGLISH AS A SECOND LANGUAGE IS STILL A PERENNIAL PROBLEM TO THE RURAL STUDENTS OF TAMILNADU." International Journal of Humanities and Social Sciences (IJHSS) 6.5, Aug - Sep 2017; 25-32

26. Ajayi, Olusola Olajide, and Olufunke Esther Ajayi. "E-Learning: A Key to Actualizing Sustainable Educational Development in Africa." International Journal of Computer Science and Engineering (IJCSE) 3.3 (2014): 109 116 (2014).