Available online at http://www.journalijdr.com

IJDR

ISSN: 2230-9926

International Journal of Development Research
Vol. 10, Issue, 10, pp. 41565-41573, October, 2020
https://doi.org/10.37118/ijdr.28840.10.2020

RESEARCH ARTICLE OPEN ACCESS

THE POWER OF SENTIMENT: BIG DATA ANALYTICS MEETS MACHINE LEARNING FOR EMOTIONAL INSIGHTS

[1]Manikanth Sarisa, [2]Venkata Nagesh Boddapati, [3]Gagan Kumar Patra, [4]Chandrababu Kuraku, [5]Siddharth Konkimalla and [6]Shravan Kumar Rajaram

[1]Sr Application Developer, Bank of America; [2]Microsoft, Support Escalation Engineer; [3]Tata Consultancy Services, Senior Solution Architect; [4]Mitaja Corportaion, Senior Solution Architect; [5]Amazon Com LLC, Network Development Engineer; [6]AT & T, Network Engineer

ARTICLE INFO

Article History:
Received 14th July, 2020
Received in revised form
03rd August, 2020
Accepted 26th September, 2020
Published online 30th October, 2020

Key Words:

Sentiment analysis, Big Data, Machine Learning, Decision Making, Support Vector Machines (SVMs).

****Corresponding author:***
Manikanth Sarisa,

ABSTRACT

The integration of Big Data and Machine Learning actually helped to transform industries throughout the global market. Among all sorts of enhancements, the one that is particularly stimulating in this field is the process of identifying emotions in large data volumes, known as Sentiment Analysis. The strength of sentiment analysis, therefore, hides inthe ability to capture the feelings, behavior and sentiments of the public, which can go a long way in helping different sectors, including marketing, politics, customer service, and health, among others.This paper also assesses the application of Big Data analytics in realizing emotional patterns supported by machine learning techniques. Using NLP on social media, reviews, and other textual data gives the business insights into the consumer's emotions, allowing them to customize their offerings better. The paper also discusses topical issues and issues encompassing these systems; noise, multiple languages and that the structures are human emotional. The paper aims to compare the machine-learning models such as the SVM, Random Forests and neural networks in their efficiency to interpret sentiments. Also, we look at the future of sentiment analysis and how it will influence the interaction between humans and machines, markets, and decision-making.

Citation: ***Manikanth Sarisa, Venkata Nagesh Boddapati, Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla and Shravan Kumar Rajaram.*** "The power of sentiment: big data analytics meets machine learning for emotional insights*", International Journal of Development Research*, 10, (10), 41565-41573.

INTRODUCTION

Today, such terms as volume, velocity and variety, along with big data in general, describe the situation when traditional means to analyze data come up short and more advanced analytical tools are required. One of the best and most useful realizations of the mentioned techniques is Sentiment Analysis. [1-3] It is the procedure of identifying the attitudes behind a text. It enables business entities and governments, among others, to know how consumers feel about the products, services or even policies they put in place; the knowledge that helps in decision-making is not found anywhere.

The Role of Big Data in Sentiment Analysis: By incorporating Big Data into sentiment analysis, the overall usage of customers' feelings and emotions in organizations has changed. It has become apparent that from a large volume of both formal and informal information, a manager can obtain substantial benefits that improve his/her decision-making. The next section will focus on how big data is used in sentiment analysis with the help of several subtopics.

- ***Volume of Data Sources:*** With the increasing emergence of big data, the amount of data produced every day has been taken to another level. A minute on a micro-blogging site such as Twitter, Facebook, Instagram and many others can produce millions of posts, comments and reviews. In contrast, online selling platforms such as Amazon, yelp,etc. contain numerous numbers of customer feedback and product reviews. It indicates that there isactually ahuge amount of data on the Internet, where organizations can analyze customer sentiments on various platforms completely. It has often been said that analytical tools are used to extract value from big data because when one processes great volumes of data, they are able to reveal various trends, patterns or anomalies that can greatly assist in making better business decisions.

- ***Variety of Data Types:*** Big data is a generalized type of information, which may be in the form of text, images, videos and audio. This variety, thus, makes it possible for the organization to capture sentiment not only from the use of texts but also from other forms of data, such as images and sounds. For instance, it means that analyzing video reviews would require identifying sentiment in spoken words, naturally provided non-verbal communication signs and movements. Sentiment analysis can reach this goal by making use of different types of data and by collecting information about the customers' emotions that are impossible or difficult to obtain using other methods.
- ***Velocity of Data Processing:*** In a world that is based on constant innovation and using technology, it is the velocity by which data is created and should be resolved, which is essential. Data streaming helps organizations get or retrieve data from social media where sentiments can be reflected in real-time to give a true reflection of what is happening in the social society. Crisis management, brand tracking and competitive assessment require this capability. For instance, in the case when negative information related to a certain brand is published, the real-time sentiment analysis enables to determine the effect of the event on the brand reputation and respond accordingly.
- ***Enhancing Data Quality and Relevance:*** Big data technologies enable the efficient carry out of more sophisticated data cleaning and data preprocessing methods thereby improving the quality and applicability of the data employed in the computation of sentiment analysis. In both these cases, as sentiment analysis is sensitive to data quality, preprocessing steps, including noise reduction, elimination of spam and low-quality text, as well as correction of misspellings, are essential. Sentiment classification can also be made more accurate through machine learning algorithms retrained to take into account other sentiment-related differences within the context of large data sets. By enhancing the standards of data, businesses can rely on the sentiment analysis result, which ultimately provides accurate results.
- ***Integration with Advanced Analytical Techniques:*** When combined with emerging technologies such as machine learning and NLP, the ability of sentiment analysis substantially improves with the help of big data capabilities. Apache Hadoop and Apache Spark framework can be used to jointly analyze large datasets in which organizations can use high-end machine learning algorithms on a large scale. Using NLP techniques, it becomes easier to capture deeper emotions from the text data resolution in richer sentiment analysis. Besides, it also aids in achieving a higher level of precisely identifying the changes in consumers' emotions between certain time intervals.
- ***Driving Predictive Analytics:*** There is no question that big data analysis is critically important to the practice of predictive analysis because it provides the basis for organizations to make predictions about future sentiment and trends based on past results. Sentiment analysis allows businesses to forecast how consumers' sentiments are likely to shift concerning the company's new products or marketing strategies or in relation to certain events. It allows commercial organizations, for instance, to predict negative issues that may come up so that you or I can get ready to resolve them before customers find out about them or positive trends, hence improving customer satisfaction and loyalty. For instance, where a firm concludes from data analysis that there is a negative attitude towards a new product launch, it can change its communication strategy to avoid a backlash.
- ***Facilitating Personalization and Customer Engagement:*** The information to be gained from big data sentiment analysis will ultimately help organizations and companies improve targeting and consumer relationships. By knowing utilitarian customer information, you can increase the efficiency of a marketing campaign, the range of offered services or products, customer relations, advertisement, and other spheres related to customers. Apart from enhancing the building of healthy relations with consumers, this sort of tactic will also improve the ability of the firm to retain current consumers and even the level of brand loyalty. For instance, in the e-commerce market, the buyer can obtain suggestions on products to be purchased depending on positive or negative sentiment analysis of a previous consumer's experience.

Volume of Data Sources
Variety of Data Types
Velocity of Data Processing
Enhancing Data Quality and Relevance
Integration with Advanced Analytical Techniques
Driving Predictive Analytics
Facilitating Personalization and Customer Engagement

Figure 1. The Role of Big Data in Sentiment Analysis

Machine Learning Techniques in Sentiment Analysis: In the current progress of sentiment analysis, machine learning approaches are typically based on the principle of categorizing textual data into three categories:happy, sad, or neutral. By using these techniques, the improvements brought in the task of sentiment analysis are way better as compared to the lexicon base method. [4,5] The current section, presents different techniques of machine learning broadly classified into different categories in a subsection.

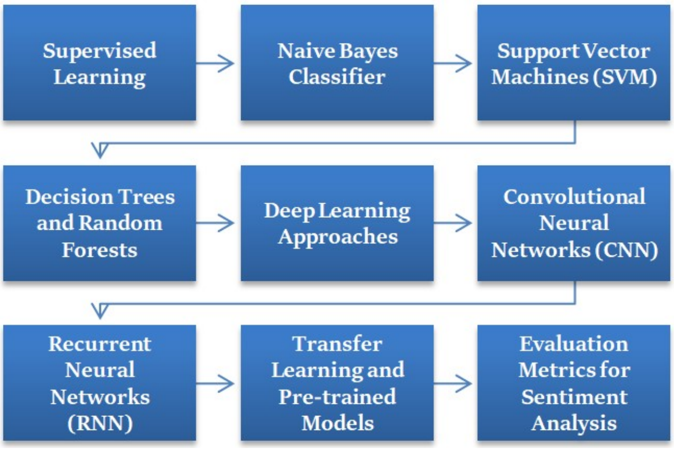

Supervised Learning
Naive Bayes Classifier
Support Vector Machines (SVM)
Decision Trees and Random Forests
Deep Learning Approaches
Convolutional Neural Networks (CNN)
Recurrent Neural Networks (RNN)
Transfer Learning and Pre-trained Models
Evaluation Metrics for Sentiment Analysis

Figure 2. Machine Learning Techniques in Sentiment Analysis

- ***Supervised Learning:*** Supervised learning is perhaps the most common form of learning used for sentiment analysis. Here, there is relabeled data where each component of the text document is assigned a sentiment (such as positive, negative, and neutral). In supervised learning used for sentiment analysis, some of the more frequently used algorithms are Naïve Bayes, Support Vector Machines (SVM) and Decision Trees. The primary strength of the use of supervised learning is its aim of being able to predict sentiments on unseen data, making it ideal when it comes to higher accuracy of the classification results. However, this approach needs a sufficiently large number of labeled data for training, which may be costly and time-consuming operations.
- ***Naive Bayes Classifier:*** A very common probabilistic classifier that can be used for sentiment analysis is known as the Naïve Bayes classifier. It is based on the fact that any feature characterizing a document, represented by a word in this case, can be present in a document independently of any other feature. Surprisingly, from the tests conducted, Naïve Bayes can perform well in text classification tasks, especially when the situation is complex and the feature space is high dimensional. Thus, it is very useful for real-time analysis but may then not be able to effectively capture the sentiment in more complex sentences in terms of the depth of different symbols and semantics of the words involved.
- ***Support Vector Machines (SVM):*** SVM is a rather stable supervised machine learning thatis applied in sentiment analysis. SVM is based on the determination of the optimum hyperplane which must best define the data respectively belonging to different categories in the higher dimensions. It means that this capability allows SVM to work well concerning complicated interrelations with features that are ideal for sentiment classification. Advantages SVM is especially useful for use in high dimensions since it leads to the separation of the classes by setting the maximum margin, hence improving the accuracy of the classifying agent. However, the tuning of parameters of SVM turns out to be arduous due to a variety of kernel functions with respect to regularization.
- ***Decision Trees and Random Forests:*** Classification trees are easy-to-understandsupervised learning models that partition data into subsets based on the featureinstances. In sentiment analysis, therefore, decision trees are able to capture non-linear relationships in the data. Random forests, which belong to the family of ensemble learning that uses decision trees, enhance accuracy in the classifier's classification by building numerous decision trees that are then combined to offer a single result. This approach reduces the influence of overfitting, hence improving the reliability of sentiment classification. While both decision trees and the random forest models are easy to implement and interpret that has made them very popular in sentiment analysis.
- ***Deep Learning Approaches:*** Recent advances in deep learning, in particular, neural networks, put such techniques as the deep learning approach on solid ground. The most frequently used technologies when it comes to sentiment classification are CNNs and RNNs. As CNNs are good at extracting local patterns and characteristics of text, they are suitable for sentiment analysis of small texts like tweets. RNNs, especially LSTMs are built for reasons of sequential data handling capacity and have a superiority to detect correlations over large distances in text to develop more profound contexts and sentiments.
- ***Convolutional Neural Networks (CNN):*** Some things we know about Convolutional Neural Networks: they are good for image processing, and the results they got in text classification tasks, including sentiment analysis are rather promising. The application of CNNs can be described by filtering the textual data locally to create features and develop spatial hierarchies in ways that teach the model valuable features. In SA, it is easy for CNNs to identify social sentiment phrases and simultaneously identify word interaction to adapt to the analysis of social media posts and product reviews. Such a feature makes the CNNs achieve almost the best accuracy in classifying the sentiment, given that they learn the hierarchical representation of data.
- ***Recurrent Neural Networks (RNN):*** Since the name RNNs suggests they are robust for sequence data, RNNs qualify to be employed for natural language processing chores such as sentiment analysis. Unlike the standard feed-forward neural networks, there are no feedback loops of information that will enable the specification of dependency duration; as a result, RNNs do possess this ability. Among all RNNs, many special types are designed for learning long dependencies in sequences and LSTM is one of the most popular networks of this kind, suitable for learning context in longer documents. Because of these features, LSTMs can be able to capture the sophistication of the sentiments expressed in the sentence thus making them highly efficient in classification of sentiment classification.

Transfer Learning and Pre-trained Models: Transfer learning needs have been increasing more in sentiment analysis with the introduction of new pre-train models like BERT and GPT. These models are trained on huge amounts of text data and can be further adjusted on particular SA tasks with great benefit from ''few-shot'' learning. The use of transfer learning enables researchers and practitioners to 'borrow' knowledge learned in other areas, which enhances the efficiency of the sentiment classification and cuts the training data set requirement.

Evaluation Metrics for Sentiment Analysis: As we know, the accuracy of the machine learning models in sense analysis decisions is imperative. Examples of such measures are accuracy, or coefficient of precision, recall rate, and F1 measure. While accuracy measures the general wrong and right aspects, precision stands for the exactness of positive sentiment. Recall checks how well the model re-pictures all the relevant positive samples, while the F1score gives equal importance to both precision and recall. Then, with the help of these metrics, a scientist will be able to understand what kind of strengths and weaknesses certain approaches to machine learning, such as sentiment analysis, have.

Literature Survey

Historical Overview of Sentiment Analysis: The term Sentiment Analysis can be traced back to the year 2000 when more emphasis was placed on extending patterns particular to text classification tasks into a new field whose focus of analysis was the emotional tones in textual data. [6-9] The lexicon based approaches formed the earliest types of models for sentiment analysis where the text was scored based on lexicon compiled over sentiment dictionaries. These dictionaries involved lists of words with corresponding sentiments in which the model was able to determine the polarity of a sentence by counting the number of positive and negative words. Despite such an approach being quite obvious and simple to apply, it had several weaknesses. In particular, the amount of contextual information and the use of idioms, irony, sarcasm and domain-specific terms affected the accuracy of the lexicon-based sentiment classification significantly. Furthermore, the practice of using dictionaries limited the ability of these models to learn about new terms and phrases and avoid using a specific word nowadays, but it was popular ten years earlier. Therefore, while initiating the process of calculating sentiment, lexicon-based methods revealed the possibility of the existence of more complex approaches being required for capturing the actual tone of the textual message.

Evolution of Machine Learning in Sentiment Analysis: Such phenomena as the Machine Learning integration into the given field can be regarded as turning points in its development. As for the methods that gained popularity with the rise of supervised learning, such as Naïve Bayes, Support Vector Machines (SVMs), and Random Forests, they became standard in sentiment classification. These models use facilitating labeled training data to identify patterns, thereby making more accurate predictions of sentiments on unseen data than the lexicon-based approaches. The capability to perform training H on a large dataset contributed to the enhancement of accuracy as well as the eligibility of sentiment analysis. It is, though as the field advanced, advanced techniques came into appraisal, especially with the appearance of Deep Learning. As for the models, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), the results also showed obvious enhancements in recognizing certain emotions. Among these, Long Short-Term Memory (LSTM) networks were popular because of the capacity within the network to preserve temporal dependencies within the sequences, which are useful for capturing context-dependent sentiments within large texts. Due to the achievements in deep learning approaches, new advancements have been made to the ways of improving the level of sentiment analysis in accepting the nuances of human language and emotions.

Applications of Sentiment Analysis: Sentiment analysis is applied across industries, and it demonstrates that it has the potential to work and deliver significant results in any field. From a marketing perspective, sentiment analysis is used to determine customer tastes and opinions in the market to enable firms to develop efficient marketing strategies, modify their products to suit the customer needs and better address the customer. From the posts on Social Media, the service reviews and the feedback, the business organizations are in a position to know the sentiments of the customers to change the campaigns. In the financial aspect, one of the most critical functions of sentiment analysis is the determination of the stock market trends based on the public sentiment in regard to some company or economic situation. This means that traders and analysts will use sentiment data in an attempt to get an understanding of market sentiment in a bid to invest successfully. Moreover, in the political context, sentiment analysis was applied to assess the voters' attitudes towards elections, which helped the political parties and candidates to adjust their strategy based on the presented attitudes. In this way, through the analysis of the recurrent psychological reactions of constituents, political campaigns might be oriented to adjust to the voters' sentimentality. In general, sentiment analysis is a useful technique for gaining insight into emotions that exist within a data set, which in turn helps to improve decision-making processes ubiquitously.

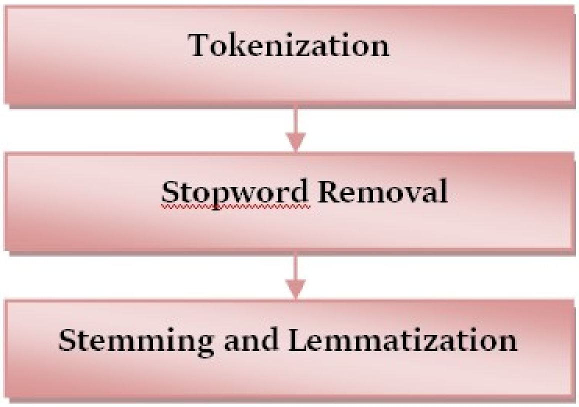
METHODOLOGY

Data Collection: This is an important step towards the achievement of sentiment analysis because the data collected is the main determinant of the quality and variation, hence the success of the analysis. The process starts with gathering massive volumes of unlabeled information from different sites that provide valuable information on the public perception.

Having a hold of the feeds from major social sites such as Twitter, Facebook, and Instagram offers real-time communal content that is perfect for customer sentiment analytics, brand insights, and tracking new industry trends. Of the two, Twitter, with its short and passionate fragments, is the best source for instant reactionsto events or products. [10-14] Multimedia posts and multimedia comments give more information about consumers' behavior and their preferences on Facebook and Instagram in comparison with Twitter. Another significant data type is the reviews that people have left on e-commerce sites such as Amazon or e-commerce, Yelp, TripAdvisor and so on, where people give their detailed appraisals regarding the product, service or experience. These are typically in the form of comments with or without ratings, which assist firms in measuring client-organizational relations and product output. Social media, blogs, new topics, and forums also provide rich information and views on issues of technology, politics and so on. For that reason, news articles, in particular, present a formal approach to identifying how people around the world feel about certain events or political topics. The data that we gather from such sources is often in the voluminous, unformatted and unprocessed state that contains high levels of noise such as idioms, emoticons, spelling mistakes and much more that is irrelevant. It is,therefore imperative to carry out a preprocessing of the data through cleaning, normalization and structuring of the data for the sake of sentiment analysis. Data preprocessing makes the data ready for consumption by machine learning models, therefore providing a way of getting insights from the emotional aspect.

Data Preprocessing: Data Under preparation is paramount to make sure that the data feeded in the machine learning models will be relevant and clean. Common preprocessing steps include:

- **Tokenization:** Some of the steps carried out on the raw text are first commonly referred to as Tokenization, as it is the first step of data preprocessing. This process assists in converting text data in a form that machine learning algorithms cannot easily interpret. In sentiment analysis, tokenization makes it easier to determine which words bear emotions, and so a more precise sentiment is derived. Tokenization is of different types: word tokenization, sentence tokenization and character tokenization. At the document level, tokenization is useful when the document is to be split into segments. In contrast, word-level tokenization is most often used in sentiment analysis, where the text is divided into Word pieces to see the sentiment of each word. Tokenization is relatively complicated in languages where there is no bound form the words, like Chinese or Japanese languages,and also where punctuation or some special characters from the structure of the special text. Tokenization is also an important aspect of sentiment analysis; proper tokenization means that important word is not loss in the preprocessing stage.



- **Stopword Removal:** Stopword removal is the process of eradicating those words that are often used in natural language and do not possess any significant importance with reference to a certain context. Such words are articles; for instance, "the" prepositions, for instance ", in" or conjunctions, for instance "and," which are omnipresent when constructing sentences but do not possess any sentiment value. While preprocessing for the use of machine learning models in sentiment analysis, It is important to exclude the stopwords since they influence the noise-to-bear ratio of actual sentiment-bearing words in a sentence. The exclusion of the most common words from the text also enhances both the speed and quality of the algorithms used to support sentiment analysis in the text. However, the problem is different in the sense that defining what exactly should be considered as stop words is completely context and language-dependent. For example, terms such as "not" can represent an extremely different message, so creating a stopword list must be done in a manner that does not eliminate critical information.
- **Stemming and Lemmatization:**Stemming and lemmatization are two procedures that are used to combine the same root words that are helpful in ordinary change of the same word. The next one is stemming,which is, in fact, much simpler than the former, whereby the final characters of each word are carved out to eliminate the prefixes or suffixes to give as words only in the stemmed form. For instance, the forms are playing, played, and plays can be put under the form of the wordplay. While stemming is prepared at quite a fast rate, it may be inconvenient for new since it can retrieve non-dictionary words or very extreme that may even alter the meaning of the word. Lemmatization, however, is a more complex technique that brings a word to the base form, which is commonly referred to as a lemma as found in the dictionary.
- Contrary to stemming, lemmmatization is a cognisant process in as much as it will only modify a word in as much as that change can be logically meaningful depending on its function in a specific sentence. For example, in lemmatization, the word 'better' is lemmatized to 'good' similar tothe word 'best', steming on the other hand, may not recognize the relationship between the two. Stemming and lemmatization thus provide important benefits in sentiment analysis models in a manner of grouping related terms together because they search for the same sentiment into different transformations of the same word in the text data.

Feature Extraction: In this phase, content data from text is converted into a numeric vector, which is easy to use in the ML algorithm. Popular feature extraction methods include:

- **Bag of Words (BoW):** Bag of Words is by far one of the simplest and most popular methods of feature extraction in text classification. It converts textual data into a set of word frequencies without respect to the arrangement of words in the data set, that is, using word bags. In this regard, a vocabulary is derived from all the documents. For each document, a vector that shows the number of times each word in the vocabulary occurs in the particular document is generated. However, BoW has several drawbacks. First, BoW is unsuitable for capturing the contextual meaning of terms because all terms are disengaged and disassociated from surrounding words. For example, in sentiment analysis, BoW has the same effect of combining words like "not good", where it will not understand that the word "not" is reversing the sentiment of the word "good." Nevertheless, BoW is very useful in cases with a large number of vectors and for models where relations between the words and their order are not of great importance. It is especially advantageous in cases where lexicon size must be restricted, and the primary concern is with the total frequency of words used.

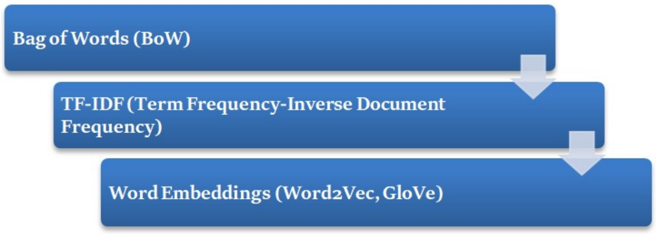

Bag of Words (BoW)
TF-IDF (Term Frequency-Inverse Document Frequency)
Word Embeddings (Word2Vec, GloVe)

Figure 4. Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF or Term Frequency–Inverse Document Frequency is a better form of feature extraction in that it takes into consideration some nonsenses of the Bag of Words model by ascertaining the manner by which the words, in a given document, are important within the document and the entire set of documents. TF-IDF thus correctly attributes high weight to words that occur commonly in the document but infrequently in the whole collection, which identifies terms useful for classifying one document against another. For instance, words like 'great' or 'awful' will be given more significance than words like 'the' or 'is' in a context like a product review dataset, quite obviously because 'the' and 'is' will not possess as much semantic relevance to an extent as 'great' and 'awful' will for a sentiment analysis context in general. In the case of the TF-IDF model, a matrix is developed from which each word is assigned a weight based on Term Frequency – the number of times that particular word occurs in a given document – and Inverse Document Frequency – how rare that word is across the entire document collection. This method is very useful in Sentiment Analysis as it allows moving the focus to sentiment-carrying words while lessening the impact of either neutral or insignificant words. Nevertheless, like BoW, the TF-IDF algorithm does not take into accountthe locality or context of words and phrases and that is why the further analysis of interrelations and distinctions between the texts is going to be challenging.

Word Embeddings (Word2Vec, GloVe): Word embeddings significantly differ from the previously discussed features, namely the bag of words and the term frequency, in that they are able to bring forth and focus on word meaning while taking into account the relationships of words as well. Accordingly, methods such as Word2Vec and GloVe are 3D or 2D 'map' coordinates for words, invariant with respect to the approximate sharing of the locations of the meanings of items, like the words. It is quite unlike the BoW and TF-IDF methods, which consider words to be detached elements from the other content, in the case of word embeddings, the word vectors gather the contextual information from where they appear in the respective corpus, making them more appropriate for example in sentiment analysis. For example, Dua (n.d.) describes the characteristics of Word2Vec, one of the most popular embedding models that run in two modes: Skip-Gram and Continuous Bag of Words. This, therefore, helps in the identification of the sentiment associated with words, for instance, if the words "happy" and "joyful" describe a particular mood, then the twain of them would be embedded together. GloVe learns in part from signal Twitter and constructs an adequate word feature using accounts from Twitter Global of Endorsers. As such, they encapsulate the essences of words and phrases while exposing them to their most delicate nuances, and these layers are critical in deep learning networks.

Classification of Sentimental Attitude: As soon as the features of the subject matter are obtained, the information is processed by sentiment classification methods. [15-18] This can either be rule-based (lexicon-based) or evaluation by machine learning approaches. The machine learning approaches may further be classified into:

Supervised Learning Algorithms: SVM, Naive Bayes, Random Forest: Pre-trained models for supervised classification are typical in sentiment analysis since the models are assigned named sentiments, positive, negative and neutral, based on training data. From these, Support Vector Machines (SVM) are commonly applied in text classification because for cut large feature spaces into zones that can classify different sentiments. SVM is most useful when the sentiment data can be made linearly separable since it yields high levels of accuracy for any number of classes of sentiment analysis. Another common supervised algorithm is Naïve Bayes, which, in fact, builds a decision system based on the Bayes theorem through independent assumption of features. Although Naïve Bayes seems to be a simple classifier, it is effective for text classification jobs because of the ease with which it handles high-dimensional data, particularly for small data sets. In contrast with the above-mentioned decision tree, Random Forest is an ensemble learning technique that creates one or multiple decision trees and fuses them to produce a more accurate and stable sentiment prediction. Not sensitive to overfitting works well with big data samples and, therefore, can successfully solve complex sentiment classification tasks. All of these algorithms performed well given specific types of input data and share the commonality of using labeled examples to identify sentiment.

Deep Learning Algorithms: CNN, RNN, LSTM: In particular, deep learning in general, and especially through the capture of dependence, that is, contextual interactions within the text, has over the years made remarkable progress towards improving the performance of sentiment classification. Convolutional Neural Networks (CNNs) have originally been designed for image classification but can also be used for text classification since here the input is considered as a sequence of word vectors. CNNs have proved to have the ability to capture local features and patterns in a text piece that includes the phrase and n-gram, and thus, they are ideal in text analysis, particularly with those containing short text like tweets and reviews. At the same time, they could have issues with the sequence of more than two-word phrases, where the sentiment is derived from even further words. For such sequential data, important structures called Recurrent Neural Networks (RNNs) are applied since they can work with sequences while saving information about words in the text inside the hidden layer. This makes the understanding of context easier by RNNs than in most ordinary models of machine learning. However, traditional forms of RNNs are difficult to work within the context of long-term dependence because of vanishing gradients. This problem is solved in Recurrent Neural Networks (RNN, especially Long Short Term Memory (LSTM)), which were designed to capture a long sequence of data streams as is the case with Reviews, Articles and Post in social media accounts. Since it has long-term memory, LSTM is one of the most beneficial algorithms to classify sentiment since often sentiments might only appear when some words are associated with other words in the sentence. They might be quite far from each other.

RESULTS AND DISCUSSION

Different approaches to sentiment analysis have brought various signs of progress in determining the emotions within big data, particularly unstructured textual information from social media networks and other online platforms, including customer reviews of products. To compare the model performances, a comparative analysis was conducted in which the performance of several established machine learning methods such as Naïve Bayes or SVM automated functions as well as deep machine learning methods such as the LSTM algorithm. The study adopted a corpus of 100,000 social media posts, which enriched the context of the experiment and offered various messages to analyze the efficiency of the models in terms of sentiment classification.

Performance of Machine Learning Models: To comprehensively evaluate the effectiveness of the various machine learning models employed for sentiment analysis, three critical metrics were utilized: These measures include; Accuracy, Precision and Recall. In addition to measures that reflect the quality of the models, these approaches employed also enable an assessment of the reliability of the models as well as their performance in real-life situations.

Accuracy: Accuracy is also one of the simplest measurements that are used to assess the effectiveness of a created sentiment analysis model. Recall, conversely, is defined as the number of true-positive and true-negative instances that occurred during evaluation to the total of all instances. In other words, accuracy can be used to say how well the model works in the context of the right sentiment recognition. High accuracy shows that,to a certain extent of the dataset, the model can correctly classify the sentiment. However, accuracy is an interesting measure, but it sometimes leads to confusion when working with unbalanced classes. For example, let's imagine 90% of the posts are labeled as positive. A model that predicted all the posts to be positive would be correct in 90% of the cases, even though it does not really keep a good distinction between positive and negative sentiments. Hence, even though accuracy is informative in a rather general way, its value should be considered hand in hand with other indicators regarding model efficiency.

Precision: Precision is more tailored to how accurately the model is at predicting all positive instances. This is defined as the Total number of accurately predicted positive samples / Total number of positive samples predicted by the models, which include true positives and false positives. To explain it in an even more basic way, accuracy tells you the proportion of the instances the model classified as positive that is genuinely positive. High precision values suggest that the fine model has a low false positive rate; therefore, it is capable of reducing the chances of wrongly predicting a negative data set as being positive. This metric is very important, especially when it incurs high penalties for false positives; for example, in sentiment analysis to brand reputation. If a brand gives a wrong classification of an aspect as positive, it can cause the formation of wrong business strategies based on wrong customer information.

Recall: True positive rate, known as sensitivity or recall rate, determines how many of the actual positive cases have been identified in the given dataset. That is, the ability of a model to correctly predict the positive instances has been defined as the proportion of correctly classified positive instances, that is, the true positive number divided by the total number of actual positive instances that are true positive + false negative. In other words, recall measures the capacity of the model to accurately locate the positive aspects that are available in the data. Upon observing high recall, it means that the model can easily identify the positive samples even at the cost of an increased number of predicted positives. This metric is particularly important in cases where it is crucial to recognize all specimens of a given sentiment. For example, in customer service instances, failure to identify a negative review (false negative) entails unmatched customer complaint dissatisfaction. However, as common knowledge, precision and recall are always in the opposite direction, meaning if one is improved, the other will be dropped; thus, the balance is crucial when creating the sentiment analysis model, which would be able to identify customers sentiments while reducing the rate of wrong sentiments noted.

Table 1. Model Performance Metrics

Model	Accuracy	Precision	Recall
Naive Bayes	78.5%	76.2%	74.1%
SVM	82.7%	80.1%	79.5%
LSTM	90.2%	88.7%	87.9%

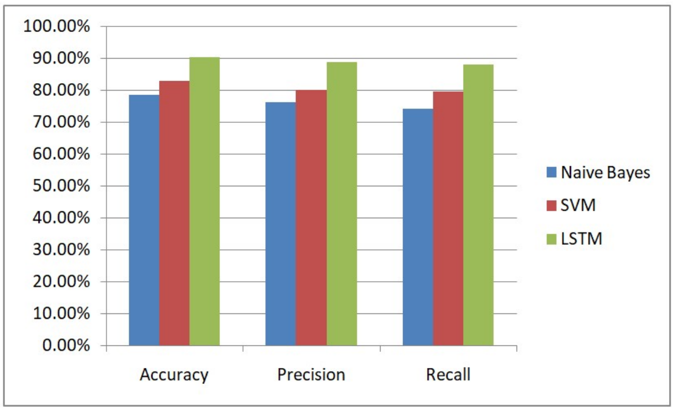


Figure 5. Model Performance Metrics

DISCUSSION

Huge advantages of LSTM over other structures are in the sentiment analysis, and they are connected with the unique architecture of LSTM, which is supposed to work with long sequences with long-range dependencies. Controlling for intonation, or in NLP, is all about the contextual and semantic dependencies between the words in the sentence to correctly gauge the sentiment in the past of text classification, techniques like Support Vector Machines (SVM), Naïve Bayes that require rather straightforward word embeddings like Bag of Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF). Although these methods are very useful in many applications, they possess some inherent weaknesses that affect their performance in sentiment analysis.

Handling Long-Range Dependencies: It can be said that one of the obvious advantages of LSTM networks is the ability to work with information from long sequences of text. This is because of their gated mechanisms, which enable the model to determine when to store or when to discard information as it makes its way through the input sequence. This capability is especially desirable in sentiment analysis because sentiments usually rely on contextual features distributed across words or phrases. For example, in a sentence like 'I like the design, but the performance is bad.' There is a positive sentiment to the design but a negative to the performance. If the model does not catch the correlation between the two extremes, it will produce the wrong overall sentiment of the message. While such context is easy for people to keep, LSTM networks are good at linking those sentiments that are manifested at different points in the text.

Recognizing Sentiment Shifts: As well, the LSTMs are good at recognizing variations of sentiment due to different linguistic features like negation and sarcasm. In NL, negation affects a statement dramatically and certainly changes the polarity. For example, "I do not like this product" is a negative sentiment message, while measuring the intensity of this sentiment and understanding the word 'this' in the context of the message may be problematic for traditional models. These shifts can be tracked effectively by LSTM networks using its memory feature so that the final classification of sentiment can reflect the right sentiment of the text.In addition to new words and phrases, sarcasm is again another problem of sentiment classification. Being able to convey a meaning opposite to what is actually said is a particularly thorny issue for models that are based solely on frequency or co-occurrence. For instance, if one you use the words like, “Great job on that project!” It might be perceived more as either praise or criticism depending on the other factors that models of the traditional paradigm may not effectively capture. On the other hand, LSTM networks are capable of correctly performing interpretation of sequences, and are thus able to determine the objective sentiment even when sarcastic or even ironical.

Suitability for Longer Texts: LSTMs are also appropriate for evaluating much longer texts; every review, article or post in social networks may contain contextual information that covers several sentences or even paragraphs. Most of the old techniques fail to handle long sequences because they lose context and, therefore, misclassify. The concept implemented in LSTMs – that of sequential processing guarantees that they are capable of identifying the patterns that reflect the overall narrative and sentiment of longer pieces of text, making them more suitable for sentiment analysis.

Error Analysis: However, some problems still exist while LSTM has higher accuracy than the other two models. Occasionally, the sentiment analysis misses sarcasm or even when the meaning and the polarity of the post are both relatively clear, the intensity of the feeling, so to speak escapes both the traditional and deep learning models alike. For instance, sarcastic text in social media can be assigned positive or even neutral connotations because of the lack of sentiment indicators. Furthermore, in cases where slang or regional expressions are used, the sentiment of the message may be misinterpreted because such words are not included in the training set

Table 2. Error Rate by Model

Model	Error Rate
Naive Bayes	21.5%
SVM	17.3%
LSTM	9.8%

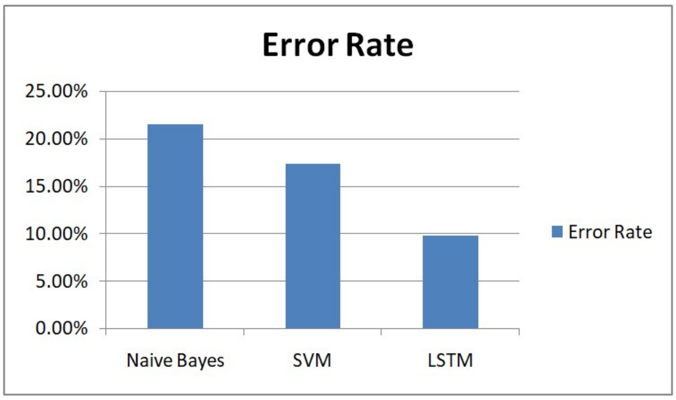


Figure 6. Error Rate by Model

CONCLUSION

The combination of Big Data Analytics and Machine Learning has employed a new dimension to the execution and stability of sentiment analysis, among various areas of industry. The availability of extensive consumer-created content like social media posts, reviews, blogs, and other web-based content is the essential advantage of which organizations can glean significant emotional data. In comparison with the previous methods of sentiment analysis, which can only provide some insights based on the texts with the help of such a limited number of features, small data sets and very simple techniques and models that are easy to implement but are not very effective in deeper analyzes of emotions and sentiments, yet with the help of big data techniques and advanced machine learning algorithms we can suddenly analyze huge volumes of sentiment information in real-time and identify intricate emotions and trends which were heretofore virtually undetectable. It gives businesses the ability to assess the customers’ views on products or services offered, manage the image of brands and, in some cases, forecast future trends in the market. The usage of different machine learning techniques and models, including simple logistic regression models with features such as Support Vector Machines (SVM) Naïve Bayes, as well as state-of-the-art deep learning models like LSTMs and CNNs, has contributed positively to the development and enhancement of the sentiment analysis techniques. These models are able to detect features of text data such as context, irony, and even shifting tonality that standard methods for analyzing such data may completely overlook. Specifically, deep learning models are extraordinary at extracting emotion from large and unstructured data and offer more detailed and complex emotions. Thus, overall confidence when making decisions based on data is higher, and this can include anything from product designing to improving services for customers or coming up with a specific advertising strategy.

In the coming years, there are a lot of opportunities that sentiment analysis will offer as ongoing developments of deep learning proceed. One of the most promising fields of development is the so-called multilingual sentiment analysis, which is an attempt to bring these types of features to different languages and cultures. Of importance, while organizations have grown multinationals, sentiment analysis in different languages has become vital for managing customer relations and the markets. Moreover, future innovations in emotion recognition will improve content analysis by extending the number of categorized emotions from positive, negative, or neutral to individual emotions like anger, joy, fear, or sadness, creating even more detailed results. Incorporating such advanced machine learning models or broad big data sources will inevitably make sentiment analysis an even more effective tool for integrating technologies that will revolutionize how organizations capture and respond to the feelings and opinions of their audience.

REFERENCES

Ahmad, M., Aftab, S., Muhammad, S. S., & Ahmad, S. 2017. Machine learning techniques for sentiment analysis: A review. *Int. J. Multidiscip. Sci. Eng,* 8(3), 27.

Alaei, A. R., Becken, S., &Stantic, B. 2019. Sentiment analysis in tourism: capitalizing on big data. *Journal of travel research*, *58*(2), 175-191.

Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. 2015. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications,* 125(3).

Bhavitha, B. K., Rodrigues, A. P., & Chiplunkar, N. N. (2017, March). Comparative study of machine learning techniques in sentimental analysis. In 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 216-221). IEEE.

Deng, L., & Yu, D. 2014. Deep learning: methods and applications. Foundations and trends® in signal processing, 7(3–4), 197-387.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Jain, A. P., & Dandannavar, P. 2016, July. Application of machine learning techniques to sentiment analysis. In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 628-632). IEEE.

Lau, R. Y. K., Zhang, W., & Xu, W. 2018. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. Production and Operations Management, 27(10), 1775-1794.

Liu, B. 2022. Sentiment analysis and opinion mining. Springer Nature.

Manning, C. D. 2008. Introduction to information retrieval.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. 2018. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32.

Mouthami, K., Devi, K. N., & Bhaskaran, V. M. 2013, February. Sentiment analysis and classification based on textual reviews. In 2013 International Conference on Information Communication and Embedded Systems (ICICES) (pp. 271-276). IEEE.

Pang, B., & Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1–2), 1-135.

Sharef, N. M., Zin, H. M., &Nadali, S. (2016). Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data. *J. Comput. Sci.*, *12*(3), 153-168.

Sharef, N. M., Zin, H. M., & Nadali, S. 2016. Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data. *J. Comput. Sci.,* 12(3), 153-168.

Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., Chung, Y. W., ... & Al-Garadi, M. A. 2018. Sentiment analysis of big data: methods, applications, and open challenges. *Ieee Access*, *6*, 37807-37827.

Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., Chung, Y. W., ... & Al-Garadi, M. A. 2018. Sentiment analysis of big data: methods, applications, and open challenges. Ieee Access, 6, 37807-37827.

Thelwall, M. 2016. Sentiment analysis for small and big data. The SAGE handbook of online research methods, 344-355.

Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. arXiv preprint cs/0212032.

Zhang, Y., & Wallace, B. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

****