International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Data Mining for Emotional Analysis of Big Data

Geda Sai Venkata Abhijith[1], Amit Kumar Vinayak Gundad[2*]

Submitted: 01/11/2022 **Accepted:** 03/02/2023

Abstract. We're in the midst of the "big data" age right now. Users generate enormous amounts of text data through a variety of means, including social media sites, e-commerce sites, and many kinds of scientific investigations. With this "Text data," companies may have a better understanding of how the public perceives their brand and use that information to guide future business decisions. As a result, it is imperative for businesses to use sentiment social media data (Big data) to generate forecasts. Open-source big data tools and machine learning techniques are needed to process massive amounts of text data in real time. To this end, we developed a machine learning algorithm-based system for analyzing sentiment in large datasets. Here, the system for text analysis system reviews datasets utilizing the Apache Spark has been built and implemented utilizing the Nave Bayes and Support Vector Machines classification techniques. In addition, accuracy was used to gauge how well the algorithms worked. As demonstrated by these experiments, the Algorithms are quite effective at managing large sentiment datasets. This will be more useful for businesses, governments, & individuals to increase their value.

Keywords: *big data analytics, sentiment analysis, and machine-learning algorithms*

1. Introduction

The Volume, Variety, And Velocity of Social Media Content Necessitates the Use of Machine Learning & Big Data Techniques to Analyze Sentiment/Text. Large And Complicated Data Collections Are Now Commonly Referred as "Big Data" (Pl. "Big Data") [1]. In Order to Derive Value from Massive Amounts of Data, New Structures and Technologies Are Needed to Capture and Analyze It [2]. Businesses And the Public At Large Benefit Greatly from Big Data. The Data Comes from A Variety of Sources, Including Sensors Meant to Collect Climate Information, social media, Video, Audio, And So On. Large Datasets Are Referred to as "Big Data" [3]. This Data can be utilized for the Sentiment Analysis [4] Since They Can Be Structured, Semi-Structured, Or Unstructured. Definition: Sentiment Analysis, Commonly Known As "Opinion Mining," Is A Method for Discovering What Authors Think About a Particular Entity.

For example, to determine whether a political party's campaign has been successful or unsuccessful, to examine movie reviews, and to examine tweets and other social media content are all examples of places where sentimental analysis is utilized [6]. Sentimental Analysis is everywhere to find out what people really think about a product, service, organization, news, movies, events, & their attributes

[4]. Apps and enterprises that monitor social media rely on the sentiment analysis & machine learning to help them learn more about brand mentions, products, and services [7].

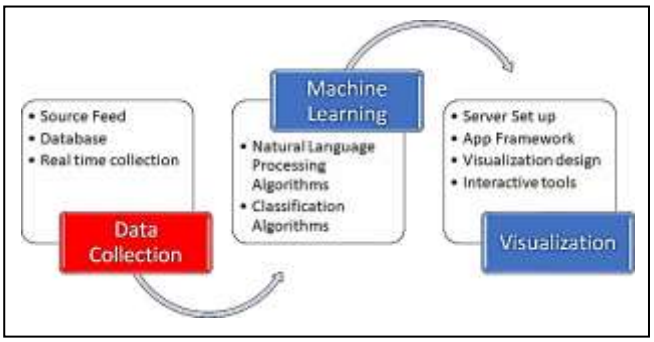


Fig.1. Example figure for sentimental analysis

[1]*Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India;*
Email ID: gedaabhijith@gmail.com; ORCID ID: 0000-0002-6740-876X
[2]*Assistant Professor, Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India;*
PhD Research Scholar, Department of Information Technology, National Institute of Technology Karnataka, Surathkal
Email ID: amit.gundad@manipal.edu; ORCID ID: 0000-0003-1820-4397
***Corresponding Author:** Amit Kumar Vinayak Gundad*
Assistant Professor, Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India.Email ID: amit.gundad@manipal.edu

The field of artificial intelligence that deals with machine learning [8] Algorithms and data are used to help computers learn and behave like humans [9]. In machine learning, a system is trained to learn from data and operate on its own based on that information. [10]. Automated learning algorithms attempt to construct predictive models from both recent and previous data. It is assumed that an algorithm would get better with practice, and this is especially true for machine learning algorithms that are trained on vast datasets and can provide impressive results in relatively confined areas [11]. A wide range of jobs in a variety of fields are assisted by machine learning. Increasingly, machine learning is being used in a wide range of applications. (Classification, Clustering, Regression, Anomaly detection, Dimensionality reduction & Recommendation) [10] are all subsets of machine learning tasks.

Apache Spark is a fast and scalable data processing engine with general application. Spark has a loop-aware scheduler and allows main-memory caching. Scala, the programming language used by Spark, runs on a Java Virtual Machine, making it easy to integrate with Java-based applications. Additionally, MLlib and Spark ML are two of Spark's machine learning libraries (also known as the Pipelines API). Because Spark SQL queries can now handle so much, operations like feature extraction and feature manipulation have never been easier. Pipeline is a more complex functionality included with Spark ML by default. Machine learning models are trained and used on data that has been cleaned, mutated, and transformed in a cycle or process before it is used for consumption and training. In the Spark ML library, a new feature called Pipeline encapsulates this full workflow of data and its stages.

2. Literature Review

Mining big data in real time [1]

A study by A. Bifet and colleagues When problems arise or new trends emerge, companies may react faster by performing real-time streaming data analysis, which allows them to discover problems and new trends as they arise and help improve their performance. The explosion of data over the last few years can be attributed to the constantly changing data sources. Every two days, we produce the same amount of data as we did from the dawn of time until 2003. For real time online prediction and analysis, evolving data streams approaches provide a low-cost, green methodology. We explore the current and future developments in data mining, as well as the obstacles that the sector faces in the coming years.

A survey on the difficulties and benefits of big data [2]

As discussed by S. Lenka Venkata et al. in their article on Big Data, referring to the data sets that are too large in order to handle utilizing the current databases' management tools, they are growing in several important applications like the Internet search & business informatics and social networks and social media. Database and data analytics studies face a monumental problem in the face of massive amounts of data. All of the amazing work being done to tackle the big data problem today. It's all about bringing people and big data together in diverse ways. Particularly, With the help of crowd intelligence, context-aware data mining, summarization and exploratory analysis of massive datasets, and privacy-preserving data sharing and analysis, we hope to present our most current work in these areas. Big data analytics systems are the focus of this article, which examines the pros and cons of each. As a way to better understand the benefits and limitations of Big Data, this study looks at several hardware platforms.

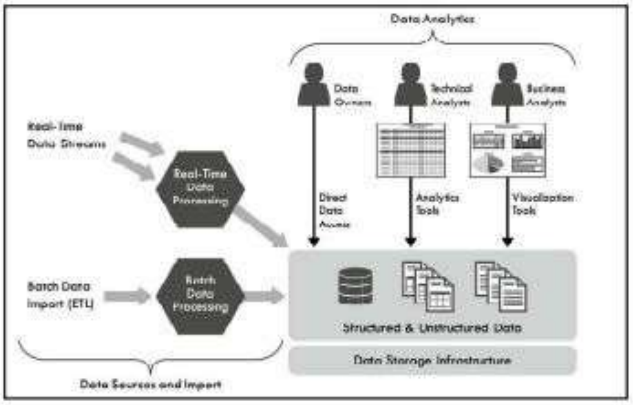


Fig.2. Bigdata architecture

Big Data Analytics: Applications & Challenges for Text, Video, Audio, and Social Media Data [3]

Machine-automated systems generate a vast amount of data in many formats such as statistical, textual, audiovisual, sensor, and bio-metric data, which is known as "Big Data," according to J. P. Verma, B. Patel, A. Smita, and P. Atul and others. Issues, problems and applications of different forms of Big Data will be discussed here with the consideration of big data dimensions in this research paper If you're interested in learning more about how to use these technologies to analyze social media and other types of information, this is the place to start. Researchers will be inspired to tackle the so-called Big Data problems of data storage, management, and retrieval. There's also a focus on India's adoption of Big Data analytics.

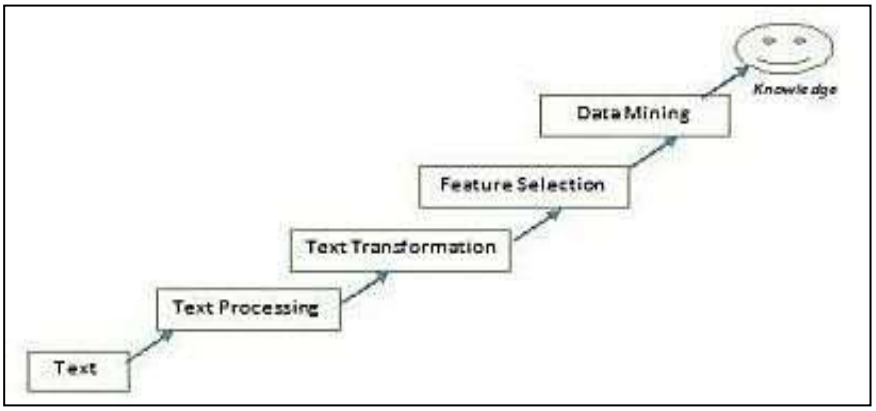


Fig.3. Text analytics system steps

Big Data Sentiment Analysis using Hadoop [4]

They, along with Merin K Kurian & Vishnuprabha V say that social media allows people a way to contact with their friends, family, and coworkers as well as a way to chat about their favorite topics (and least favorite brands). Unstructured dialogue can give organizations important insight into how the consumers view their brand & allow them to make certain business decisions to protect the image of their brand. Researchers are becoming increasingly interested in Sentimental Analysis and Opinion Mining as the number of the sentiment-rich social media onto the web grows rapidly. Sentiment Analysis, on the other hand, is now considered a Big Data problem because of the abundance of social media. The primary goal of the study was to discover a method for performing Sentiment Analysis on large datasets that was both effective and efficient. A huge dataset of tweets was subjected to Sentiment Analysis in this study, and the speed and accuracy of the method were evaluated. For large sentiment datasets, it appears that the technique is extremely efficient.

Support Vector Machine Sentiment Analysis [5]

The orientation of a sentence is classified as either negative or positive using sentiment analysis, which is considered as classification task. In order to develop a sentiment classifier, we used support vector machines (SVMs) on benchmark datasets. These features were extracted using N-grams and a different weighting scheme. Using Chi-Square weights, it identifies the most useful features for the classification process. Using Chi-Square feature selection may improve classification accuracy, according to the results of an experiment.

A Framework for a Cloud-Based Hybrid Recommender System for Big Data Mining

Modern e-commerce websites that offer a plethora of products use recommendation systems that automatically suggest new, intriguing items to users, thereby retaining them according to Kamal Al-Barznji, Atanas Atanassov et al. Recommendation systems aid users in their information management. In this paper, a structure for Cloud - delivered Hybrid Recommender System (FCHRS) for Big Data Analysis is proposed, along with a discussion of the framework's techniques. The system also incorporates Sentiment Analysis (also recognized as opinion mining) as well as the Iterative collaborative altering technique, which is the most common method. In order to discover and extract subjective information from source materials, this technique makes use of text analysis, natural language processing, & computational linguistics. Because of this, the firm must mine social media (big data) to come up with ideas for new products and services. By combining the outputs of two algorithms, it is possible to obtain more actionable data. This is a brand-new field of study that has yet to be explored.

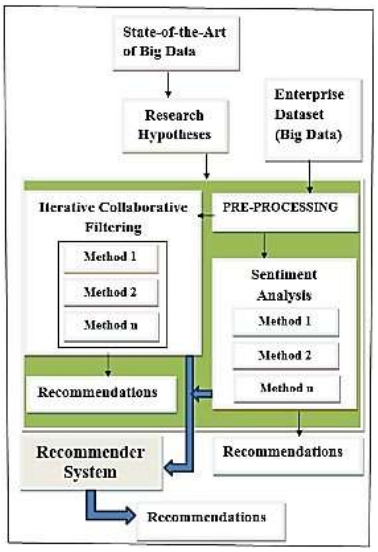


Fig.4. Big Data Mining for Generating Recommendations

3. Implementation

There are a number of techniques used in Sentiment Analysis to discover and extract subjective information from texts and other source sources [7]. The internet is a valuable source of information on people's feelings. Users can upload their own material on a variety of social media platforms, including forums, microblogs, and social networking sites [14]. Analyzing online reviews for sentiment is a way to get a fuller picture of what people have to say about a product. Because it categorizes the content of a text as either positive or negative, sentiment analysis is still considered a classification task. Sentiment classification is frequently accomplished through the application of machine learning. Using sentiment analysis, a variety of research, including reviews of consumer products and services, has been analysed [5]. Sentimental analysis is one of the types of text analysis that can be used. The sentimental analysis technique can be used in case you possess a piece of text & want to know what type of emotion it evokes (e.g., love vs. hate, positive vs. negative). To this end, we developed a machine learning algorithm-based system for analyzing sentiment in large datasets.

The text classification methods were devised and developed here. Naive Bayes & SVM classification algorithms from supervised machine learning algorithms are utilized to do sentiment analysis on selected data sets in this system, which uses Spark ML library to create this system. In order to evaluate the offered procedures, Accuracy was used to gather data. Figure 5 depicts the structure of the sentiment analysis system used to make predictions from large datasets. Predictive models that can tell if a sentence is favorable or negative are the purpose of this project.

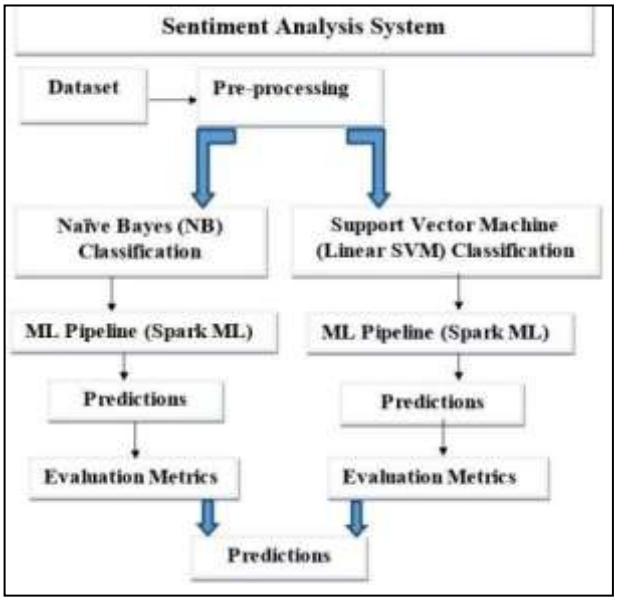


Fig.5. Sentiment Analysis Algorithms for analysing massive amounts of data and making predictions.

Steps

To begin, the dataset is imported into the system as input, and sparse data processing is then applied.

Following data input into the system, the dataset is split into 80 percent training and 20 percent testing datasets at random.

Once the training dataset has been pre-processed, sentiment analysis will begin with techniques such as Feature Extractors and Feature Transformers, both of which are NLP concepts. As long as the code for both algorithms is essentially identical, just the model utilized for Nave Bayes and Linear SVM models will change.

The remaining steps in this operation are nearly identical for the two algorithms, all the way through to the end.

Computer-assisted translation

Some natural language processing (NLP) concepts and techniques will be employed to solve sentimental analysis challenges, such as the following:

An observation's attribute or property is represented by a feature. It's also known as a parameter. A feature, on the other hand, is a stand-alone variable. There are rows for each observation and columns for each feature in a table dataset. When it comes to user profiles, for example, consider a tabular dataset that includes fields such as a person's age, gender, occupation, city, and so on.

As a tool for text mining, a feature vectorization technique known as Term Frequency-Inverse Document Frequency (TF-IDF) is frequently employed. However, TF and IDF have been separated in MLlib to allow for greater flexibility. Term frequency vectors can be generated with either HashingTF or CountVectorizer. HashingTF is a transformer that transforms sets of terms into feature vectors of fixed length. A "set of terms" in text processing can be a collection of words. An IDFModel is created by fitting an Estimator to a dataset. In the IDFModel, feature vectors (produced via HashingTF or CountVectorizer) are scaled.

Words in a Bag (BoW). Bag-of-words is a text representation that shows where specific words appear in a document. In this approach, each word is treated as a separate feature. It is necessary to turn text into numbers before using machine learning algorithms because the algorithms cannot deal with raw text information directly. A vector is exactly a set of numbers.

With tokenization, an input string (text) is transformed to lowercase and separated into words utilising whitespaces as a separator. This is possible with a basic Tokenizer class. Word sequences are used to segment sentences.

Stop words are words that ought to be excluded from the input because they occur frequently but convey little meaning. Stop Words Remover accepts as input a string sequence (the output of a Tokenizer) & eliminates all stop words from input sequences.

Algorithms

Data is used to train a model in machine learning techniques. Fitting a model to data is a synonym for training a model. A distinction can be made between supervised and unsupervised machine learning algorithms based on the type of training data.

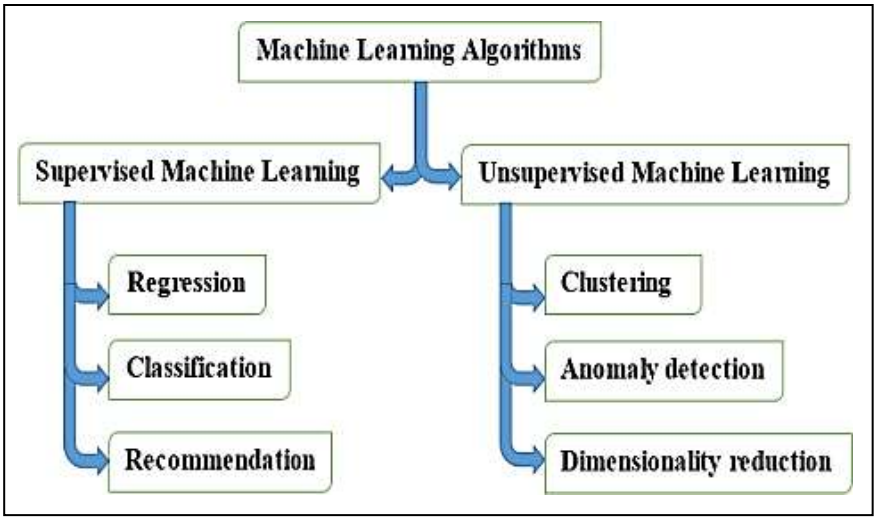


Fig.6. Machine learning algorithms

Supervised Machine Learning Algorithms

The term "supervised machine learning" refers to the process of learning using pre-labeled data sets [8]. They can only be used with training datasets that have been tagged. In the training dataset, each observation has its own attributes and a label. One or more predictor variables

are used as predictors in supervised machine learning algorithms, which learn from the data (features). In a training dataset, the labels can either be created manually or obtained from another system. Recommendation engines, regression models, and classification models are three subcategories of supervised machine learning algorithms [12]. (Decision trees, Linear regression, & Ensembles of trees), for classification problems include: (SVM, Logistic Regression, Naive Bayes & Neural Network), & for recommendation tasks include the Collaborative filtering using ALS [10].

Algorithms for Unsupervised Machine Learning

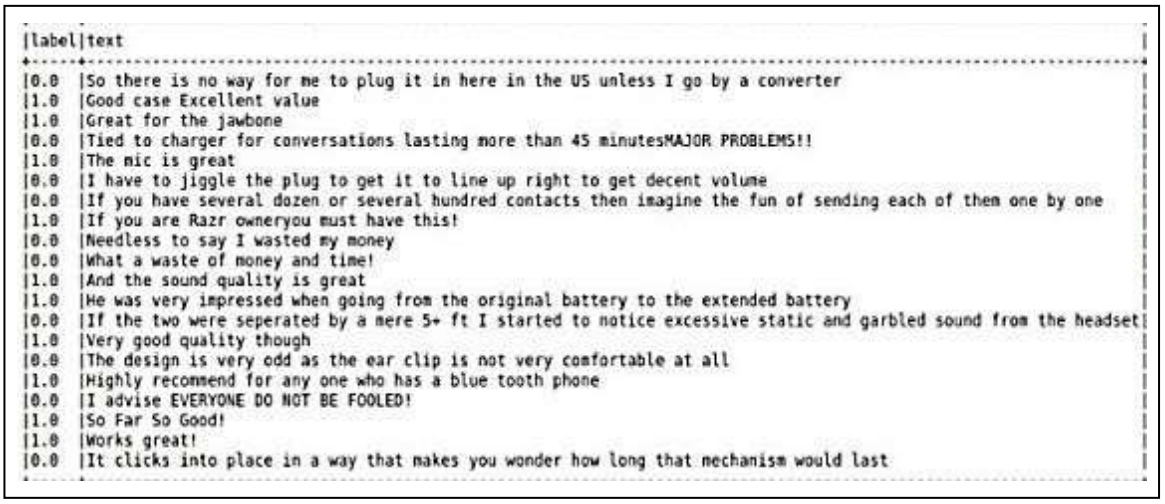
It is utilized when the dataset is not being labeled, which means that the model do not require labelled data which is called unsupervised learning Unlabeled data can be used as a source of information for these models, which are designed to learn or identify hidden structures in the data. There is no right or wrong answer in unsupervised learning [8]. Anomaly detection, clustering, and dimensionality reduction are all common applications for them. Singular value decomposition (SVD) and principal component analysis (PCA) are two extensively used unsupervised methods [10].

When it comes to classification, Bayesian Classification is a statistical and supervised learning method. Naive Bayes Classification Algorithm It is capable of diagnosing and forecasting issues. Based on the Bayes theorem, Naive Bayes is a simple technique for classifying multiclass data. 'The Bayes theorem is founded on a experience learning concept, which is employing a series of steps to arrive at a prediction. An event's probability can be calculated based on prior knowledge of possible causes [6]. The likelihood that a data point belongs to a particular class is computed using the Nave Bayes model, which is a probabilistic one [12].

Support Vector Machine (SVM) is among the most frequently used supervised machine learning algorithms. This approach is employed in several real-world applications, including text categorization and image classification. Regression and classification can both benefit from SVM, which is why it is so often used. Instead of predicting classes according to whether or not the model is evaluated positively or negatively, it56 differs from Nave Bayes in this regard [12]. Spark supports a binary classifier implementation for linear SVM.

4. Experimental Analysis

For supervised machine learning algorithms in sentiment analysis, this dataset is pre-labeled and ready to go. There are datasets of sentiment labelled sentences available at https://archive.ics.uci.edu/ml. It was produced for the paper [17], this dataset has a sample of reviews from the three websites, amazon cell phone products reviews (amazon.com), movies reviews (imdb.com), & restaurant reviews (yelp.com). It has 500 favorable and 500 negative reviews that were chosen at random from each website. A bad review is denoted by the number 0, whereas a favorable review is denoted by the number 1. The tab character separates a review from its label. However, for this technique, only Amazon.com reviews were used. The file amazon cell labelled.txt contains these customer testimonials.



label	text
0.0	So there is no way for me to plug it in here in the US unless I go by a converter
1.0	Good case Excellent value
1.0	Great for the jawbone
0.0	Tied to charger for conversations lasting more than 45 minutesMAJOR PROBLEMS!!
1.0	The mic is great
0.0	I have to jiggle the plug to get it to line up right to get decent volume
0.0	If you have several dozen or several hundred contacts then imagine the fun of sending each of them one by one
1.0	If you are Razr owneryou must have this!
0.0	Needless to say I wasted my money
0.0	What a waste of money and time!
1.0	And the sound quality is great
1.0	He was very impressed when going from the original battery to the extended battery
0.0	If the two were seperated by a mere 5+ ft I started to notice excessive static and garbled sound from the headset
1.0	Very good quality though
0.0	The design is very odd as the ear clip is not very comfortable at all
1.0	Highly recommend for any one who has a blue tooth phone
0.0	I advise EVERYONE DO NOT BE FOOLED!
1.0	So Far So Good!
1.0	Works great!
0.0	It clicks into place in a way that makes you wonder how long that mechanism would last

Fig.7. Amazon dataset

Train and test datasets

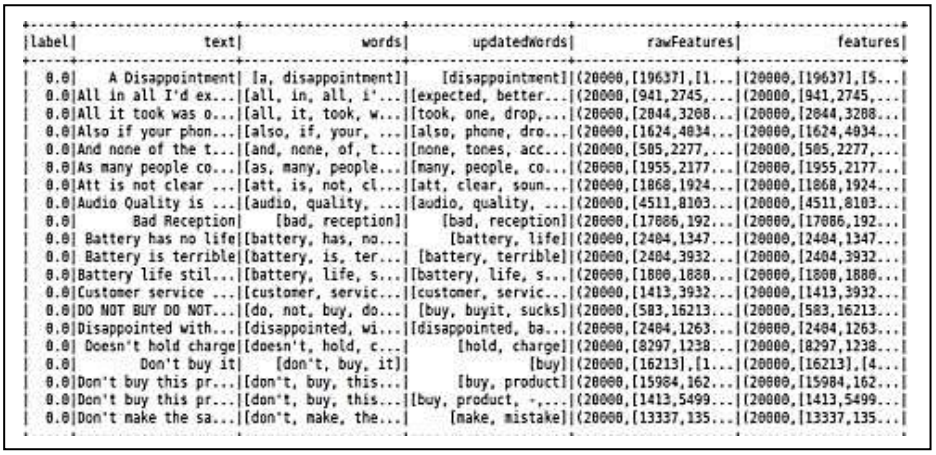
The datasets will be divided into two categories: training data and testing data, in order to calculate the generalization error [9]. Before implementing and assessing models, the datasets are randomly divided into two categories: 80 percent train datasets and 20 percent test datasets. As a result, the term "training dataset" describes a collection of data utilized in model development for the purpose of developing predictions and recommendations from the models. However, the test

dataset is a separate set of data that is not utilized in the creation of model. The test dataset is used to evaluate the quality of the training models by finding performance indicators such as accuracy.

Data that has been pre-processed

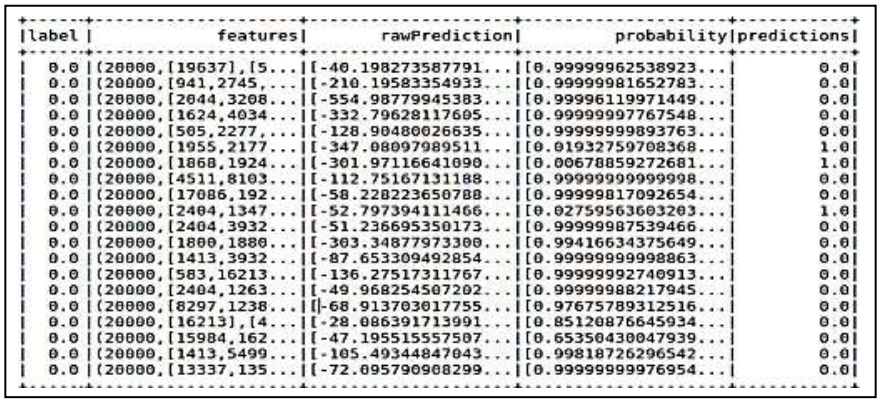
As soon as the dataset is loaded, it is randomly split into 80 percent training (80% of the data set) and 20 percent testing (the remaining 20%). On top of that, we used NLP concepts (Features Transformers and Extractors) to clean up and pre-process the train dataset. We used whitespaces to separate words (tokens) and then removed those that weren't needed. Until the data is ready for machine learning algorithms, it cannot be used. The TF-IDF method is used to generate a bag of words for each sentence in the dataset, hence a feature vector must be created for each one. Sentimental analysis methods were then applied using a word bag

.



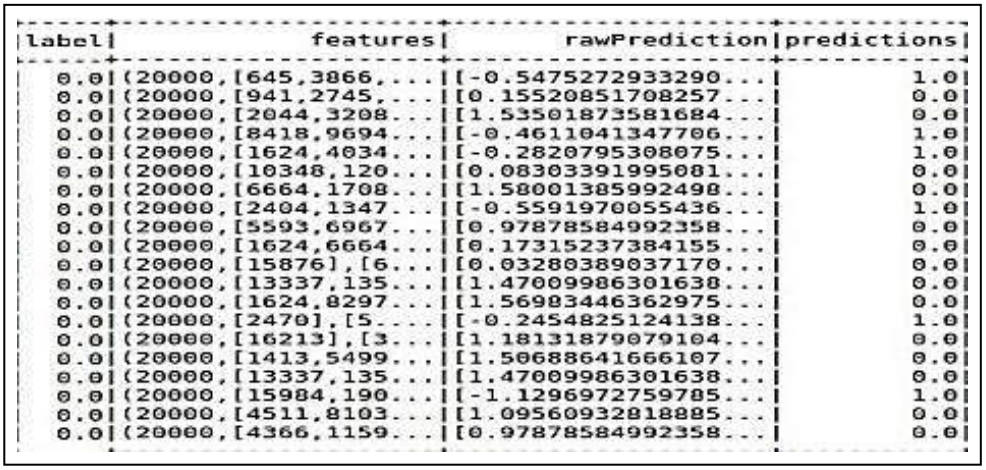
label	text	words	updatedWords	rawFeatures	features
0.0	A Disappointment	[a, disappointment]	[disappointment]	(20000,[19637],[1...	(20000,[19637],[5...
0.0	All in all I'd ex...	[all, in, all, i'...	[expected, better...	(20000,[941,2745,...	(20000,[941,2745,...
0.0	All it took was o...	[all, it, took, w...	[took, one, drop,...	(20000,[2044,3208...	(20000,[2044,3208...
0.0	Also if your phon...	[also, if, your, ...	[also, phone, dro...	(20000,[1624,4034...	(20000,[1624,4034...
0.0	And none of the t...	[and, none, of, t...	[none, tones, acc...	(20000,[505,2277,...	(20000,[505,2277,...
0.0	As many people co...	[as, many, people...	[many, people, co...	(20000,[1955,2177...	(20000,[1955,2177...
0.0	Att is not clear ...	[att, is, not, cl...	[att, clear, soun...	(20000,[1868,1924...	(20000,[1868,1924...
0.0	Audio Quality is ...	[audio, quality, ...	[audio, quality, ...	(20000,[4511,8103...	(20000,[4511,8103...
0.0	Bad Reception	[bad, reception]	[bad, reception]	(20000,[17086,192...	(20000,[17086,192...
0.0	Battery has no life	[battery, has, no...	[battery, life]	(20000,[2404,1347...	(20000,[2404,1347...
0.0	Battery is terrible	[battery, is, ter...	[battery, terrible]	(20000,[2404,3932...	(20000,[2404,3932...
0.0	Battery life stil...	[battery, life, s...	[battery, life, s...	(20000,[1800,1880...	(20000,[1800,1880...
0.0	Customer service ...	[customer, servic...	[customer, servic...	(20000,[1413,3932...	(20000,[1413,3932...
0.0	DO NOT BUY DO NOT...	[do, not, buy, do...	[buy, buyit, sucks]	(20000,[583,16213...	(20000,[583,16213...
0.0	Disappointed with...	[disappointed, wi...	[disappointed, ba...	(20000,[2404,1263...	(20000,[2404,1263...
0.0	Doesn't hold charge	[doesn't, hold, c...	[hold, charge]	(20000,[8297,1238...	(20000,[8297,1238...
0.0	Don't buy it	[don't, buy, it]	[buy]	(20000,[16213],[1...	(20000,[16213],[4...
0.0	Don't buy this pr...	[don't, buy, this...	[buy, product]	(20000,[15984,162...	(20000,[15984,162...
0.0	Don't buy this pr...	[don't, buy, this...	[buy, product, -,...	(20000,[1413,5499...	(20000,[1413,5499...
0.0	Don't make the sa...	[don't, make, the...	[make, mistake]	(20000,[13337,135...	(20000,[13337,135...

Fig.8. Preprocessed data



label	features	rawPrediction	probability	predictions
0.0	(20000,[19637],[5...	[-40.198273587791...	[0.99999962538923...	0.0
0.0	(20000,[941,2745,...	[-210.19583354933...	[0.99999981652783...	0.0
0.0	(20000,[2044,3208...	[-554.98779945383...	[0.99996119971449...	0.0
0.0	(20000,[1624,4034...	[-332.79628117605...	[0.99999997767548...	0.0
0.0	(20000,[505,2277,...	[-128.90480026635...	[0.99999999893763...	0.0
0.0	(20000,[1955,2177...	[-347.08097989511...	[0.01932759708368...	1.0
0.0	(20000,[1868,1924...	[-301.97116641090...	[0.00678859272681...	1.0
0.0	(20000,[4511,8103...	[-112.75167131188...	[0.99999999999998...	0.0
0.0	(20000,[17086,192...	[-58.228223650788...	[0.99999817092654...	0.0
0.0	(20000,[2404,1347...	[-52.797394111466...	[0.02759563603203...	1.0
0.0	(20000,[2404,3932...	[-51.236695350173...	[0.99999987539466...	0.0
0.0	(20000,[1800,1880...	[-303.34877973300...	[0.99416634375649...	0.0
0.0	(20000,[1413,3932...	[-87.653309492854...	[0.99999999998863...	0.0
0.0	(20000,[583,16213...	[-136.27517311767...	[0.99999992740913...	0.0
0.0	(20000,[2404,1263...	[-49.968254507202...	[0.99999988217945...	0.0
0.0	(20000,[8297,1238...	[-68.913703017755...	[0.97675789312516...	0.0
0.0	(20000,[16213],[4...	[-28.086391713991...	[0.85120876645934...	0.0
0.0	(20000,[15984,162...	[-47.195515557507...	[0.65350430047939...	0.0
0.0	(20000,[1413,5499...	[-105.49344847043...	[0.99818726296542...	0.0
0.0	(20000,[13337,135...	[-72.095790908299...	[0.99999999976954...	0.0

Fig.9. Sentiment analysis by using Naïve bayes



label	features	rawPrediction	predictions
0.0	(20000,[645,3866,...	[-0.5475272933290...	1.0
0.0	(20000,[941,2745,...	[0.15520851708257...	0.0
0.0	(20000,[2044,3208...	[1.5350187358168 4...	0.0
0.0	(20000,[8418,9694...	[-0.4611041347706...	1.0
0.0	(20000,[1624,4034...	[-0.2820795308075...	1.0
0.0	(20000,[10348,120...	[0.08303391995081...	0.0
0.0	(20000,[6664,1708...	[1.58001385992498...	0.0
0.0	(20000,[2404,1347...	[-0.5591970055436...	1.0
0.0	(20000,[5593,6967...	[0.97878584992358...	0.0
0.0	(20000,[1624,6664...	[0.17315237384155...	0.0
0.0	(20000,[15876],[6...	[0.03280389037170...	0.0
0.0	(20000,[13337,135...	[1.47009986301638...	0.0
0.0	(20000,[1624,8297...	[1.56983446362975...	0.0
0.0	(20000,[2470],[5....	[-0.2454825124138...	1.0
0.0	(20000,[16213],[3...	[1.18131879079104...	0.0
0.0	(20000,[1413,5499...	[1.50688641666107...	0.0
0.0	(20000,[13337,135...	[1.47009986301638...	0.0
0.0	(20000,[15984,190...	[-1.1296972759785...	1.0
0.0	(20000,[4511,8103...	[1.09560932818885...	0.0
0.0	(20000,[4366,1159...	[0.97878584992358...	0.0

Fig.10. Sentiment analysis by using linear SVM

The amazon cell labelled dataset was used to test both machine learning techniques for text analysis. In the end, after training models on a train dataset, the results were tested against the test dataset. Using only the characteristics and their associated probabilities and predictions, the Nave Bayes model generates these results. The LinearSVM model, on the other hand, just generates predictions and can be easily compared to the labels' base values.

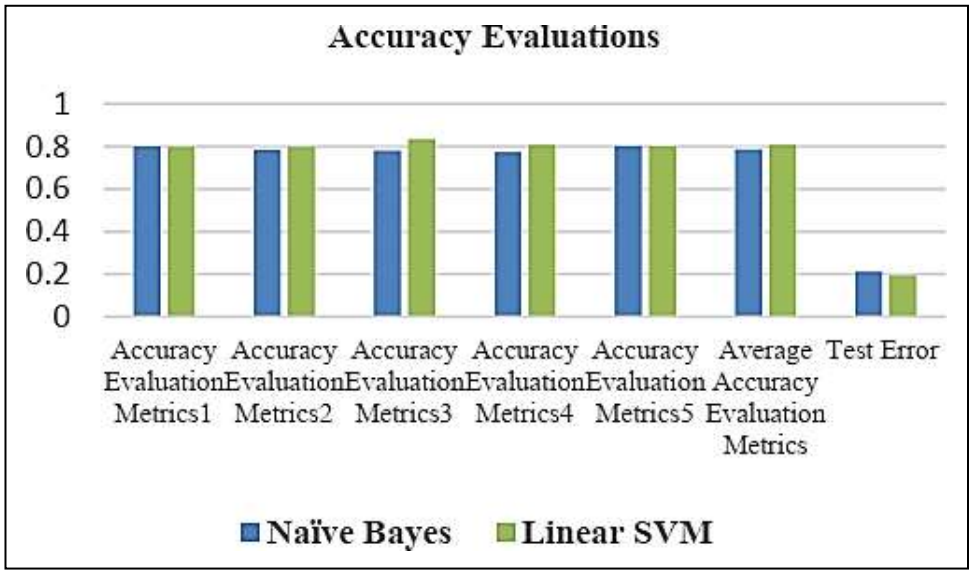


Fig.11. Evaluation metrics for both algorithms

The LinearSVM model outperforms the Nave Bayes model in terms of average accuracy assessment metrics, based on the test datasets used to evaluate this dataset. In other words, if the accuracy score is greater than or equal to 1, then there is less error. supervised machine learning techniques and big data machine learning libraries are used for all of the above operations.

5. Conclusion

It's one of the most intriguing approaches for gauging consumer sentiment about a product. In this study, we propose and test a system that can conduct near-real-time sentiment analysis on massive amounts of data moving at a rapid rate. Big data and large-scale data processing are the result of our utilization of Apache Spark. There was a positive and negative sentiment classification in the proposed approaches (Nave Bayes and SVM). The proposed models have passed through the preprocessing stage, feature generation stage, classifiers learning stage, and Pipeline step. A model's correctness is measured as part of the evaluation process. A cross-validation approach to improving classification precision has been demonstrated through experiments to improve the system's overall performance by getting average accuracy metrics for evaluations. In terms of average accuracy metrics, Linear SVM outperforms Nave Bayes in the suggested system's evaluation analysis.

6. Future Scope

The primary goal of this study was to speed up Sentiment Analysis such that large data sets could be processed more effectively. This work has been done onto a single node, & while it is expected to perform better within a multi-node enterprise-level configuration, it will need to be tested in the future with much larger data sets to verify that it does indeed perform better.

References

[1] A. Bifet, "Mining big data in real time," Inform., vol. 37, no.1, pp.15–20, 2013.

[2] S. Lenka Venkata, "A Survey on Challenges and Advantages in Big Data," vol. 8491, pp. 115–119, 2015.

[3] J. P. Verma., A. Smita, B. Patel., and P. Atul, "Big Data Analytics: Challenges and Applications for Text, Audio, Video, and Social Media Data," Int. J. Soft Comput. Artif. Intell. Appl., vol. 5, no. 1, pp. 41–51, 2016.

[4] Ramesh R, Divya G, Divya D, Merin K Kurian, and Vishnuprabha V, "Big Data Sentiment Analysis using Hadoop", IJIRST, Volume 1, Issue 11, pp. 92-98, 2015.

[5] Nurulhuda Zainuddin, Ali Selamat," Sentiment Analysis Using Support Vector Machine", IEEE International Conference on Computer,

Communication, and Control Technology (I4CT 2014), Kedah, Malaysia, pp.333-337, 2014.

[6] Rajat Mehta,"Big Data Analytics with Java", Published by Packt Publishing Ltd, ISBN 978-78728-898-0, UK, 2017.

[7] Kamal Al-Barznji, Atanas Atanassov, "A Framework for Cloud Based Hybrid Recommender System for Big Data Mining", a journal of "Science, Engineering & Education", Volume 2, Issue 1, UCTM, Sofia, Bulgaria, pp. 58-65, 2017.

[8] Jason Bell, "Machine Learning: Hands-On for Developers and Technical Professionals", Published by John Wiley & Sons, Inc., Indianapolis, Indiana, 2015.

[9] Boštjan Kaluža, "Machine Learning in Java", first published: Published by Packt Publishing Ltd, UK, 2016.

[10] Mohammed Guller, "Big Data Analytics with Spark", ISBN- 13 (pbk): 978-1-4842-0965-3, 2015.

[11] Benjamin Bengfort and Jenny Kim, "Data Analytics with Hadoop", Published by O'Reilly Media, Inc., First Edition. USA, 2016.

[12] Nick Pentreath, "Machine Learning with Spark", Published by Packt Publishing Ltd. BIRMINGHAM – MUMBAI, 2015.

[13] https://spark.apache.org/docs/latest/, Online Feb 2018.

[14] X. Fang and J. Zhan, "Sentiment analysis using product review data," J. Big Data, pp. 1–14, 2015.

[15] https://machinelearningmastery.com/Online Feb 2018.

[16] Alexandros Baltas, Andreas Kanavos, and Athanasios K. Tsakalidis, "An Apache Spark Implementation for Sentiment Analysis on Twitter Data", Patras, Greece, Springer International Publishing AG, pp. 15-25, 2017.

[17] Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth, "From Group to Individual Labels Using Deep Features", Publication rights licensed to ACM, KDD, Sydney, NSW, Australia, 10 pages, 2015.