JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(1), MARCH, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.net

Machine Learning Based Sentiment Analysis Using Natural Language Processing

Tofunlope Abijo, Ken Mcgarry
Faculty of Technology,
School of Computer Science,
University of Sunderland, United Kingdom.

ABSTRACT
The said approach uses multiple statistical and machine learning based methods to mine the hidden data patterns in comprehensive and diversified datasets. In other words, artificial intelligence is a backbone of data mining, as it gives the ability of making autonomous decisions using a particular dataset. We have performed the process of classification using multiple AI based mining techniques such as Random Forest, Decision Tree and Linear Regression classifiers.

ARTICLE INFO
Article History
Received: August, 2023
Received in revised form: August, 2023
Accepted: January, 2024
Published online: February, 2024

KEYWORDS
Sentiment Analysis; Classification; Data Mining; Random Forest; Decision Tree; Regression.

INTRODUCTION

Due to recent advancement in the domain of information communication technology, there exists a cluster of deep knowledge from which a user intends to retrieve relevant knowledge. In order to retrieve this knowledge, a user has to perform certain operations such as data mining. The Data mining has emerged as a power tool for solving the analytical problems such as decision making which enables a particular organization to gain a competitive advantage in a corporate sector.

The data mining algorithms can be implemented to extract the desirous information from a large data repository such as tweets from Twitter. Using a platform of social media such as Twitter, we can extract relevant content. As it contains updated information about a particular topic and gives insights on-going trends. For example, “so we rescued a pup... meet rexy” now this tweet attracts attentions of users towards animal rescue. Most of population all across the globe uses twitter to express their thoughts and suggestion suggestions. Twitter provides an instant communication to its users within a text limit of 140 characters (Feldman, 2017).

A tweet is a textual communication which contains views of an individual about a particular object, person or a product. It has been observed that twitter also contains multiple negative tweets which indicates un-satisfactory behavior of a particular object. Similarly, the powerful tool of data mining can be employed to perform specific data operations such as predictive analysis, text analysis and sentiment analysis etc. Sentiment Analysis (SA) is a contextual mining of text which finds and extracts textual information in a relevant source material, and helping an organization to evaluate the social sentiment of a particular brand, product or service while observing word of mouth on multiple social media platforms. However, a sentiment analysis of a social media comments is generally restricted to a basic sentiment analysis and count based metrics (Medhat, 2018).

With a recent development in the domain of deep learning, the capability of a learning algorithm to analyze a particular text has been improved significantly. Smart use of advanced learning techniques could be an effective tool for doing in-depth research. We believe that it is quintessential to classify customer’s communication on following key points such as:

1. To identify key aspects of a product and services that customers care about.
2. Evaluating customers’ intentions and reactions concerning those aspects.

These basic concepts when used in a combination, becomes a vital tool for analyzing millions of human conversations with human level accuracy (Prabowo, 2019). Machine learning is an efficient approach which can be employed to perform a particular mining task (Sajda, 2018) such as tweet classification based on sentiment analysis. In scientific literature, most of the earlier studies have addressed this issue, however, most of them are found to be

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

complaining about non-availability of a proper dataset (Ashima Yadav, 2019). Furthermore, it is a general requirement of almost every learning approach that input data should be large and diversified. In order to address aforementioned issue, there is a need of a comprehensive and diversified dataset which should contain relevant sentiment tags such as negative, positive or neutral. Therefore, to conduct this study a comprehensive and diversified corpus of tweets have been collected. The selection of this corpus is two-fold as it contains a diversified and comprehensive dataset and it is available publicly. Moreover, this study emphasizes on mining and classification of tweets as positive or negative using a domain of data mining and analytics. Furthermore, to implement machine learning based mining methodologies three techniques are employed such as Decision Tree, Linear Regression and Random Forest using given testing and training datasets.

Statement of Problem

In light of recent advancements in information communication technology, there has been a proliferation of deep knowledge repositories from which users seek to extract pertinent information. To accomplish this, users must undertake various operations, including data mining, which has emerged as a potent tool for resolving analytical challenges such as decision-making, thus empowering organizations to gain a competitive edge in the corporate sector.

Data mining algorithms offer the means to extract desired information from vast data repositories, such as Twitter. Leveraging social media platforms like Twitter allows for the extraction of relevant content, offering insights into ongoing trends and updated information on specific topics. For instance, tweets serve as textual communications reflecting individual views on various subjects, including objects, persons, or products. However, alongside positive sentiments, Twitter also hosts negative tweets, indicative of unsatisfactory experiences or perceptions. While data mining facilitates specific operations like predictive analysis, text analysis, and sentiment analysis, the latter, although valuable, is often constrained to basic sentiment analysis and count-based metrics.

Nevertheless, recent developments in deep learning have significantly enhanced learning algorithms' capacity to analyze text, presenting an opportunity for more in-depth research. Recognizing the importance of classifying customer communications, it is imperative to identify key aspects of products and services that customers prioritize, as well as evaluating their intentions and reactions regarding these aspects. This holistic approach serves as a vital tool for analyzing millions of human conversations with human-level accuracy.

Machine learning emerges as an efficient approach for performing mining tasks such as tweet classification based on sentiment analysis. While prior studies have addressed this issue, many have lamented the lack of an adequate dataset. Consequently, there is a pressing need for a comprehensive and diversified dataset containing relevant sentiment tags, such as negative, positive, or neutral.

To address this need, this study focuses on collecting a comprehensive and diversified corpus of tweets for mining and classification purposes. Utilizing machine learning-based methodologies, including Decision Tree, Linear Regression, and Random Forest, the study aims to classify tweets as positive or negative based on sentiment analysis.

In summary, this study endeavors to contribute to the field of data mining and analytics by providing insights into sentiment analysis of tweets and exploring machine learning-based methodologies for tweet classification.

Objectives of the Study

This research work is based on machine learning based sentiment analysis using natural language processing. The following sets of research objective are suggested.

1. To investigate the efficacy of data mining algorithms in extracting relevant information from large data repositories, with a focus on tweets from Twitter, for decision-making processes in organizations.
2. To explore the application of sentiment analysis techniques, particularly in the context of social media platforms like

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

Twitter, for evaluating the social sentiment of brands, products, or services.
3. To assess the impact of deep learning advancements on the capability of learning algorithms to analyze textual data, and to determine the effectiveness of advanced learning techniques in conducting in-depth research on customer communication patterns.
4. To develop and implement machine learning-based methodologies, including Decision Tree, Linear Regression, and Random Forest, for the mining and classification of tweets as positive or negative sentiments, utilizing comprehensive and diversified datasets publicly available.

LITERATURE REVIEW

This study emphasizes on two categories such as sentiment analysis and machine learning. The first category of this study is sentiment analysis where sentiment analysis is commonly employed for text mining like tweets, new and brand reviews etc. It is enormously implemented in the Natural Language Processing (NLP), linguistic computations, and evaluation of a particular text, however, it significantly helps in identifying, extraction and quantification of a subjective information (Li, et al., 2017). The major objective of sentiment analysis is to work widely in a form of a word of mouth such as reviews or responses on any product. For instance, a customer wants to order an online product item like apparel, grocery or any other home appliances, so before ordering that item a consumer will review most of the opinions, comments and complaints about that item or product on a relevant or social media platform. Similarly, it will help pivotally to make strategic and accurate decision about that item (Tatemura, 2018) (Liu, 2017).

In literature sentiment analysis is defined using three terms. Such as, a product about which a review is placed, product features, most important is consumer who is giving review about that product. Sentiment Analysis encompasses multiple boundaries such as accurate identification of a relevant product, feature extraction and validity of word of mouth. The classification based on Sentiment Analysis can be implemented using multiple steps such:

1. Classification in terms of document
2. Classification using sentences
3. Classification based on features or aspects

Arrangement at a report level is carried out to where one needs to order by and large polarization of a theme regardless of a shopper (Agarwal, 2020). In record level-opinion examination a survey about single article is introduced in a specific report. It tends to be considered as evident, for instance if there should be an occurrence of a verbal, film ubiquity, and so forth, where an archive contains remarks about a solitary film or a solitary item. The sentences are considered as a crucial piece of a specific language in any case, it is a more limited type of archive as an assortment of sentence relates to one report (Liu, 2017). In a Sentence-level-order, each sentence is considered as a solitary remark. This order can be classified two sub-classifications: item recognizable proof and audit location. Likewise, at an element level or viewpoint level, the examination on different provisions of an item is performed. Assume a buyer willing to buy a Samsung Mobile Phone, then, at that point above all else he will unequivocally underline on item elements, for example, the camera nature of the PDA is reasonable yet the sound nature of the cell isn't reasonable. So In request to assess different parts of an item, the perspective level assessment is by and large performed (Lin, 2018).

The opinion investigation incorporates different advances, for example, Data Pre-preparing, determination of pertinent provisions and exact order to assess the class of an important information. The Data pre-pre-processing can be additionally sorted after advances like tokenization, expulsion of stop words and so forth Tokenization incorporates separation of sentences into an array of word such as the tokens. Stop words incorporates following words, for example, am, are, in, to, the, at etc. therefore they don't contain any assessment. Consequently eliminate them prior to utilizing them in an important examination. Additionally, stemming is an errand of changing over a specific word into its root structure such 'combination' can be changed over into root structure 'solidify' (Whitelaw, 2019).

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

Second area of this study is employment of machine learning techniques. Machine learning approach is one of the prominent approaches which can used to address multiple problems (Sajda, 2018) such as sentiment analysis. Machine learning is a domain of Artificial Intelligence which enables a machine to learn autonomously and improve its capabilities from its environment without being programmed explicitly. Machine Learning emphasis on development of computer programs that can learn and access multiple resources on their own. The learning algorithm observes input source data, such as earlier examples, direct experience, or instruction, in order to find out patterns in data and make a robust decision making in the future on a relevant input test data. The primary aim of machine learning is to make intelligent machines that can learn autonomously without human involvement or assistance. Likewise, the usage of a classical machine learning algorithms in text is mostly employed for keyword sequences; instead, a technique based on semantic analysis involves ability of a human being to understand meaning of a text. There are multiple types of learning algorithms such as supervised, non-supervised, and semi-supervised and reinforcement learning algorithms (Liu, 2020).

In supervised machine learning algorithms, algorithms can follow what has been discovered in the previous to new facts the usage of labeled examples to predict future events. Starting from the evaluation of a recognized education dataset, the studying algorithm produces an inferred feature to make predictions about the output values. The device is in a position to grant pursuits for any new enter after enough training. The gaining knowledge of algorithm can additionally examine its output with the correct, supposed output and locate blunders in order to regulate the mannequin accordingly. In unsupervised machine learning algorithms, algorithms are used when the facts used to learn is neither labeled nor labeled. Unsupervised learning research how structures can infer a characteristic to describe a hidden shape from unlabeled data. The system doesn't determine out the proper output, however it explores the records and can draw inferences from datasets to describe hidden buildings from unlabeled data (Gonçalves, 2019).

In semi-supervised machine learning algorithms, algorithms fall somewhere in between supervised and unsupervised learning, due to the fact they use each labeled and unlabeled information for coaching – commonly a small quantity of labeled records and a giant amount of unlabeled data. The structures that use this approach are in a position to significantly enhance gaining knowledge of accuracy. Usually, semi-supervised studying is chosen when the received labeled facts requires expert and applicable sources in order to train it / learn from it. Otherwise, obtaining unlabeled information commonly doesn't require extra resources. The reinforcement machine learning algorithms is a learning method that cooperates with its environment by producing activities and discovers faults or rewards. Hit and trial approach is used in these types of algorithms which cause delayed output, are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to autonomously determine the best performance within a specific context in order to achieve optimal performance. Simple feedback is required for a machine to learn which action is best; this is known as the reinforcement signal (Taboada, 2019).

In other words, Machine mastering permits evaluation of big portions of data. While it normally gives you faster, greater correct effects in order to discover worthwhile possibilities or risky risks, it may additionally require extra time and assets to educate it properly. Combining laptop studying with AI and cognitive applied sciences can make it even greater fine in processing massive volumes of information. In scientific literature most of the studies have addressed issue of sentiment analysis using machine learning approaches such as Support Vector Machine, Neural Network, Decision Tree, Rule based classifier etc.

2.1. Support Vector Machine

SVM was at first created to resolve issues of twofold characterization. Its significant target is to focuse on deciding the best grounds which can go about as separator to expand choice limits between different datapoints from

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

different classes. The premise ought to be chosen which ought to keep a most extreme separation between at least two help vectors of various classes. The SVM holds ability to execute both straight, and non-direct characterization errands (Zadeh, 2017).

In this category (Zainuddin & Selamat, 2019) an article employed SVM as a classifier with multiple averaging approaches like TF-IDF, the total abundance of most frequent terms, Binary Occurrence etc. For selection of best features, they employed a renowned technique which is which is commonly used for subtraction of useless features and noise removal from a given dataset. According to study authors, the proposed technique works significantly with SVM to produce better results.

In like manner, another examination introduced a clever method for feeling investigation and assessment mining utilizing SVM with a higher exactness of 0.78 as far as accuracy. SVM can be classified into managed learning method dependent on its dynamic. As per the relationship of various items, classes can isolate choice limits. The sentence or tweet recovered from a twitter is can be characterized into different classes like positive, negative or nonpartisan dependent on SVM preparing calculation. There are four SVM models. Two of them are utilized for order and other can be utilized for relapse SVM. SVM methods can be prepared to limit blunder work utilizing: E ¼ 1=2 WTE þ C Xbi ð4þ, where C is steady and b is a boundary for non-distinguishable information input. W is a vector coefficient. SVM executes a part capacity to plot input highlight vector space into another vector space where classes are directly detachable. This methodology utilizes a polynomial bit which is firmly reliant upon store size, type, resilience and numFolds.

The feature-vectors from multiple articles like nouns, verbs, adverbs and adjectives are made from their input coefficient extracted from predefined grades of their sentiments. The values lesser than 0.5 corresponds to a negative influence whereas values greater than 0.5 leads to positive sentiment and coefficients with 0.5 value are meant for neutral class. (Prastyo, 2020)

2.2. Artificial Neural Network

Artificial Neural Network (ANN) is persuaded development a goliath supernatural occurrence of nature as it impersonates plan of a human frontal cortex. The significant unit of a neural connection is neuron. ANN wires a data layer, stowed away layer, and a yield layer. A section vector is given as an assurance to a neuron, where vector infers the repeat of a word in a record. There is a weight "A", what separates to each neuron which is done to work out cost work. Neural alliance relies on a speedy cutoff is x (I) = Aa. The sign of x (I) is used to portray a class.

In ANN based organizing of a program contained two stages: forward and in back causing. In forward spread, the information is given at the data layer of neurons which is imitated by the stacks which are intrepid numbers. Cutoff centers are utilized to standardize the yield respect some spot in the level of 0 and 1. Then, at that point the yield is disconnected and the objective worth, suffering there is a capacity between two credits then, at that point in alter spread is performed. During backpropagation input is copied just so end up in regards to so that weight can be changed. As such learning relies on mishandle. (Sachdeva and Sabharwal, 2018) used a neural relationship for face gathering which has given a high precision rate.

(Vega and Mendez-Vazquez, 2018) proposed a Dynamic Neural Network (DNN) Model where he utilized continuing on and Hebbian learning for the learning structure. He separated the model system and DNN and showed that DNN acts in a maintained manner over benchmark strategies. Patil et al. (Patil, 2017) proposed a technique where he utilized latent semantic assessment (LSA) with a convolution neural association (CNN). LSA is a technique for changing word over to vector. Weighting in LSA performed with TF-IDF assessment. His model gives 87% precision. Another appraisal performed evaluation depiction of online audits utilizing BPN. BPN is an adaptable learning methodology with a limit of assortment of sentiments from social information. For each status plan, the information sources are applied to the affiliation. Neurons at focus points of first layers are first thing orchestrated with loads of depicted reach. Then, at that point at stowed away layer particular mixes of bigrams and tri-grams showing up in sentences are thought of. With the livelihoods of these moving mixes, loads get

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

changed while going to the yield layer. Utilization of fumble work signs to select weight changes. As indicated by number of things, action words and modifiers present in sentences, number of neurons in three layers shifts.

Su et al. presented Ensemble of learning for opinion arrangement in 2017. They have used Chinese Lexical Semantics to join estimates of various base models. Mix of base models with reconsidering of planning data by calling base classifiers. Naturally drawing sub instances of planning data and subsequently by joining larger part projected a polling form classes will give most ideal assumption for gathering.

The proposed work develops a Treebank for chines sensations of social data to vanquish the absence of checked and gigantic corpus in existing models. To predict the imprints at sentence level i.e positive or negative, the Recursive Neural Deep Model (RNDM) was proposed and achieved prevalent than SVM, Nave Bayes and Maximum Entropy. 2270 film reviews were accumulated from the site and Chinese word division gadget ICTCLAS was used to piece these studies. Five classes were made due with each sentence and standford parser applied for sentence parsing. The proposed model chipped away at the assumption for assessment names of sentences by wrapping up 13550 chines sentences and 14964 words. ME and NB performs higher with contrastive mix structure than baselines with unbelievable edge.

In another examination, the maker has proposed a model for assessment examination pondering both visual and printed substance of casual associations. This new arrangement used significant neural association model, for instance, Denoising auto encoders and skip gram. The establishment of the arrangement was CBOW (Continuous Bag-Of-Words) model. The proposed model involved two segments CBOW-LR (key backslide) for text based substance and was reached out as the CBOW-DA-LR. The gathering was done by the limit of visual and printed information. Four datasets were surveyed, i.e., Sanders Corpus, Sentiment140, SemEval2013 and SentiBank Twitter dataset. The proposed model outmaneuvered the CBOWS+SVM and FSLM (totally managed probabilistic language model). Perhaps the ESLAM (extended totally regulated probabilistic language model) in term of little planning data had beated the current model. The component learning and skip grams both required gigantic datasets for best execution. (Nti, 2019).

2.3. Decision Tree

It is a tree-like approach where the non-terminal focuses address a section and terminal spotlight point looks out for the name. The way is standard the explanation of a condition. This is a recursive correspondence at last appear at a terminal neighborhood which gives a carving to an information.

The fundamental test in the choice tree is to see which quality is to be picked as a root place point. This can be settled by utilizing some quantifiable methodology, for example, data gain and Gini report. A choice tree is an enchanting method for assessment evaluation since it moreover gives a decent outcome on a gigantic stack of information. Typically utilized choice tree assessments are CART, CHAID, and C5.0.

A tree pulls out the getting sorted out information persistently. For the division of information, a condition is utilized that is on the brand name worth. Condition subject to whether a word is open or missing. The division cycle is proceeded until terminal focus places address the little extents of parts which are utilized for the mentioning task. Kotenko et al. (Kotenko, 2017) utilized a choice tree to hinder the joke substance on the site. He utilized TF-IDF for weighting the word which tells about the significance of the word and a binomial classifier which brief concerning if a word has a spot with a particular outline.

2.4. Naïve Bayes

The Naïve Bayes classifier treat each word as free so it can't find a semantic connection between the words in any case Bayesian Network can. Bayesian connection unequivocally considers the words dependence on each other. The Bayesian alliance watches out for dependence in term of an organized outline which is non-cyclic where each center point keeps an eye out for the word as a variable and edges address the dependence between the parts. As a tendency classifier, Al-Smadi et al. (Al-Smadi, 2018) utilized Bayesian

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

affiliations, discovering certifiable yield and every so often high, when stood apart from different classifiers.

An examination on-assessment mining utilizing Naïve Bayes classifier and illuminating mix taken from Manhattan Hierarchical Cluster in 2013. NB classifier model is an arranged non-cyclic framework whose spotlights pass on parts and edges contains restrictive conditions. In message procedure for feeling assessment BN obviously is rich.

2.5. Nearest Neighbor

An appraisal has inspected KNN while separating SVM and different classifiers. It utilizes three kinds of distance works explicitly Euclidean, Manhattan and Minkowski for discovering opening between two terms under get-together measure. In this cycle a case is portrayed by utilizing most likeliness of its lining respects, the case being appropriated to the class with among its K closest neighbors perceived by one of above distances.

2.6. Decision Tree

Jaskarn and Shveta in 2019 presented "Analysis and recognizing verification of Human Emotions through Data Mining". It is a reformist based classifier gives separating of preparation information space where worth of some allocating is utilized to seclude information. Division of information things or explanations if there should arise an occasion of text mining is done recursively so that last leaf communities contain tokens for get-together.

The author have proposed the arrangement of huge learning for appraisal of twitter. The rule point of assembly of this work was to introduce the heaviness of cutoff points of convolutional neural affiliation and it is crucial to set up the model unequivocally while keeping away from the fundamental of adding new part. A neural language is utilized to instate the word embeddings and is prepared by huge execution social event of tweets. For additional refining the implanting on colossal composed corpus, a standard neural affiliation is utilized. To instate the relationship, actually implanted words and cutoff points were utilized, having same planning and preparing on coordinated corpus as of Semeval2015. The parts utilized in proposed work are incitations, sentence cross segment pooling, softmax and convolutional layers. To set up the affiliation, stochastic propensity plunge (SGD) and non-raised cutoff improvement calculations were utilized and to find the focuses back affecting assessment was utilized. Dropout technique were utilized to redesign the neural affiliations regularization. The huge learning model is applied on two undertakings: message level assignment and explanation level errand from Semeval-2015 as far as possible and accomplish high results. By applying six test-set, the proposed model lies at first circumstance in a truly significant time-frame of accuracy.

In the wake of exploring this stack of studies, it is set up that by utilizing huge learning procedures, evaluation can be created in really convincing and unmistakable manner. As the appraisal is utilized to anticipate the perspectives on clients and huge learning models are about the guess or copy of human frontal cortex, so the huge learning models give more accuracy than shallow models.

RESEARCH METHODOLOGY

To address the raised issues in this study, we present a comprehensive methodology such as collection of a benchmark dataset, data pre-processing, implementation of data mining classifiers such as Random Forest, J48 Decision Tree Classifier, Naïve Bayes etc.

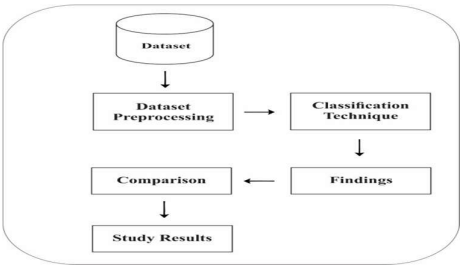

Figure 1: Research methodology diagram.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

3.1. Benchmark Dataset

In order to evaluate a particular mining approach there is a need of a benchmark dataset which should contain all type of relevant information. For this purpose, we collected a dataset from Analytics Vidhya which contains fully mapped dataset. The selection of this dataset is two-fold as it is publicly available and it contains a diversified dataset of 31,962 tweets.

3.2. Data Attributes

In this dataset, there are two type of attributes such as label and Tweet:

3.2.1. Tweet

Twitter is a trending medium used by the wider population to express their thoughts, from which some of the tweets draws the attention of audience towards the content of these tweets. Twitter is a social media platform, which allows its users to communicate their thoughts instantly with the public over the web in the form of tweets limiting to the 140 characters. The follower can like, share and make a comment on a particular tweet. Twitter also allows its users to create tags in their tweets indicated by the '#' symbol, which is also known as the hashtag. The hashtag represents specific information like #birthday, #iPhoneX etc. Based on these hashtags, twitter recommends the tweets based on its users' interests. This functionality makes it a popular social media platform than its counterparts [4].

3.2.2. Label

A label indicates a category of a instance, such as in case of tweets we can classify tweets into two categories such as Positive Tweets and Negative Tweets etc.

3.3. Data Preprocessing

In a next step, a vigorous analysis is performed in order to find out any missing values or other data issues. It can be observed from figure 1 below that there are no missing values.

```
df.isnull().any()

id       False
label    False
tweet    False
dtype: bool
```

Figure 2: Analysis of missing values.

3.3.1. Count of Unique Values

Figure 1 below presents a count of unique class values. It can be observed from a chart below that it contains a count of 29,720 values as a 0 and 2,242 as 1. Similarly, a class value 0 indicates a negative tweet and 1 can referred as positive.

Table 1: Count of unique values.

Sr. #	Encoded Value	Class
1.	0	Negative
2.	1	Positive

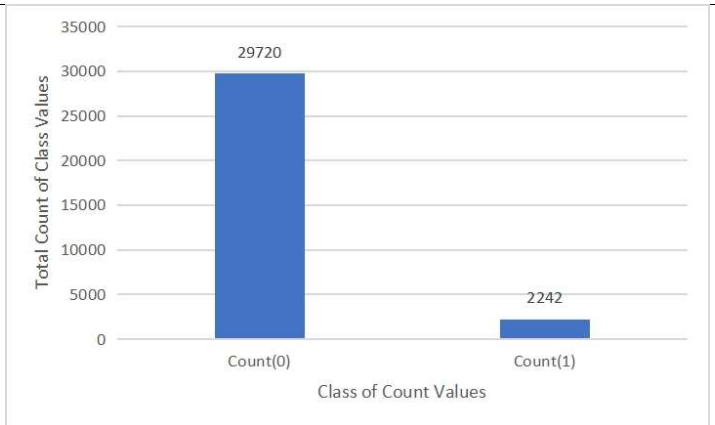

Figure 3: Count of Class Values.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

3.3.2. Count of Punctuation Values

Punctuation is a usage of symbols such as " ", ' ', "!" and spacing etc. to create a meaning full sentence. It is general requirement of a sentiment analysis that a data should be properly cleaned in terms of punctuation marks as they create a noisy symbols in a particular data. Therefore, it is mandatory to remove all punctuation marks from a particular text in order to perform specific tasks such as sentiment analysis or text classification etc. Figure 3 – 5 below draw an illustration of a removed punctuation marks from an input dataset of tweets.

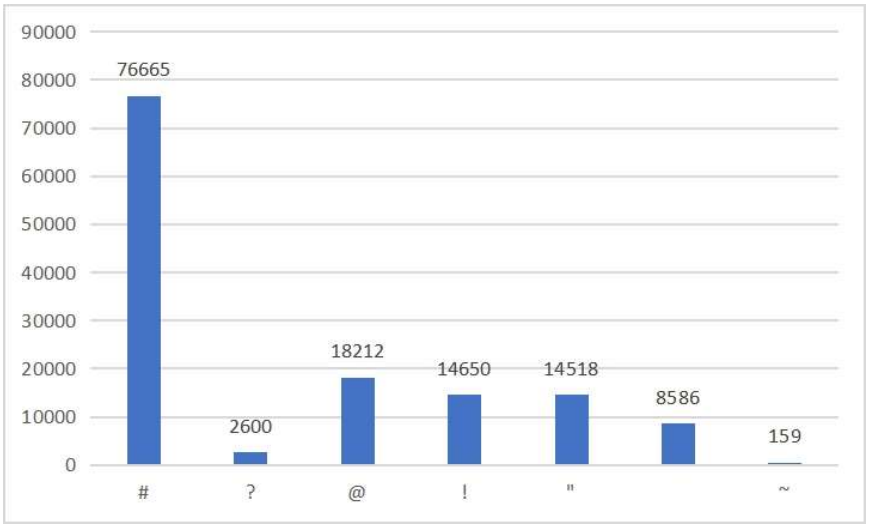

Figure 4: Count of Punctuation Marks.

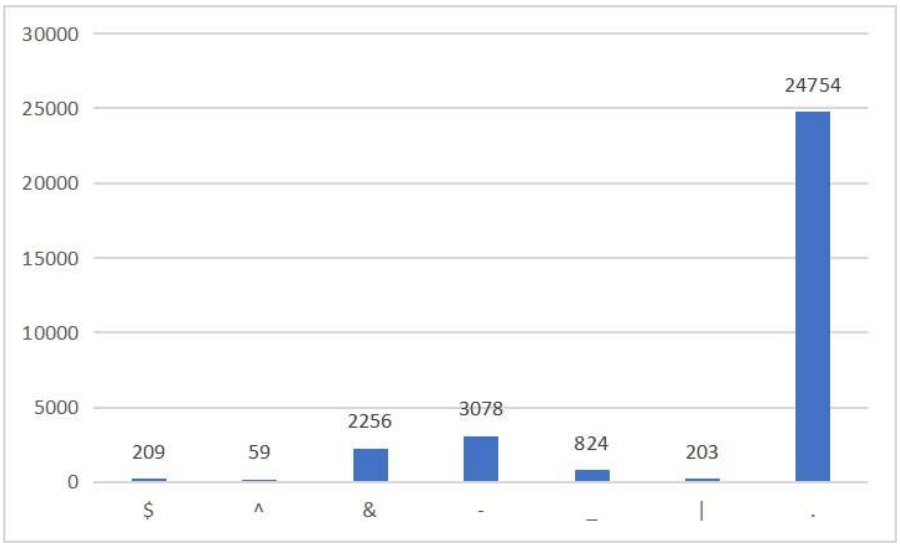

Figure 5: Count of Punctuation Marks.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

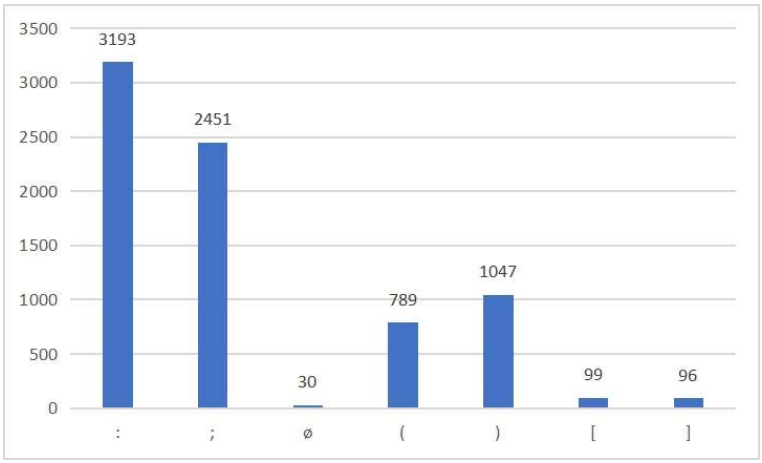

Figure 6: Count of Punctuation Marks.

3.3.3. Demographical Description

A figure 6 below presents a demographical description of obtained dataset in terms of standard deviation, mean, minimum value, maximum value, count etc. A standard deviation is statistical measure which is used to find out a diversity in a given dataset. This dataset has reported a standard deviation of 9226.78 which means that input data is diverse in nature. Mathematically, this equation can be written as:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \quad \ldots\ldots (1)$$

Similarly, a mean value is considered as an average value which is a weighted sum of all elements in a particular dataset whereas a value of 15981 is obtained as a mean value from this dataset. Mathematically, it can be represented using a particular equation such as:

$$m = \frac{\text{sum of the terms}}{\text{number of terms}} \quad .. (2)$$

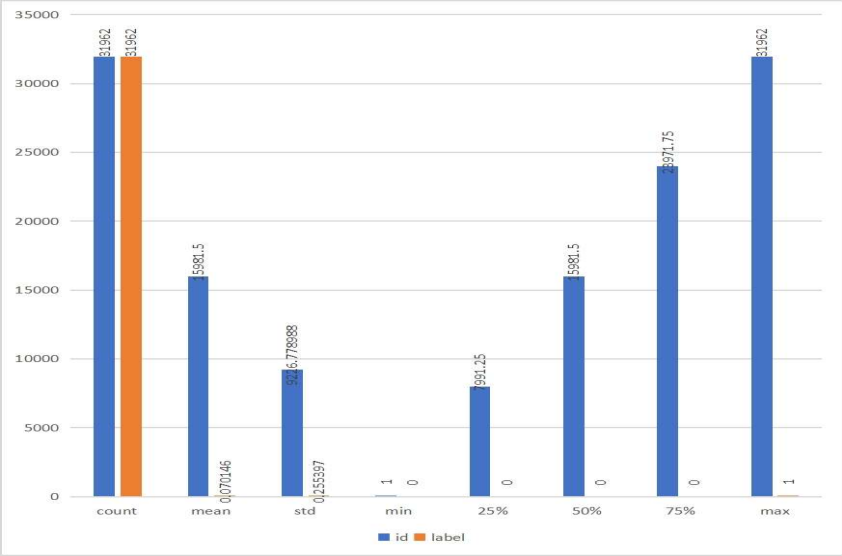

Figure 7: Demographical Description of a Data.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

Machine Learning Classifiers

Machine learning is an efficient approach which can be employed to perform a particular mining task (Sajda, 2018) such as tweet classification based on sentiment analysis. Therefore, in this phase, we implemented machine learning approach using different methodologies such as Decision Tree, Random Forest and Regression for the experimentation.

3.3.4. Decision Tree

Decision tree-based machine learning approach is an effective approach in terms of the supervised learning. A decision tree is a kind of approach which uses a tree alike model which contains multiple decisions with their possible outcomes. The normal tree contains a parent node, child nodes, branches, and leaves. In a decision tree, terminal nodes contain the test cases and normal terminal nodes have decision outputs (Bogumił Kamiński, 2017).

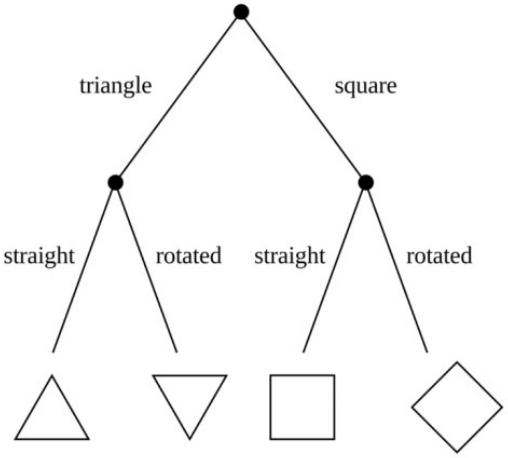

Figure 8: Example of a Decision Tree.

3.3.5. Random Forest

Random forest, is an effective ensemble based supervised learning technique. Usually, a random forest consists over two or more individual decision trees that operate as an ensemble. However, each individual tree in the random forest outputs a class prediction and the class having a higher number of votes will be used for prediction (Ho, 2010).

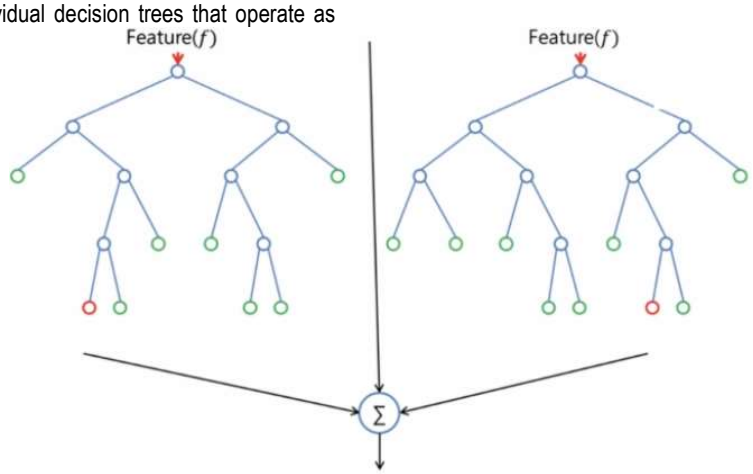

Figure 9: Example of a Random Forest.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

3.3.6. Regression

Regression is a supervised machine learning approach which is used for a prediction of target which should be continuous in nature and produces a constant slope. Generally, it predicts values within a particular range such as 0 – 1 etc (Freedman, 2009).

RESULT DISCUSSION, PROJECT MANAGEMENT AND RELEVANCE

Results and Evaluation

As discussed earlier, three machine learning based mining approaches are employed such as Decision Tree, Random Forest and Regression. In this experimentation, separate datasets were employed for testing and training purpose. Furthermore, in order to evaluate these techniques three evaluation are selected such as precision, recall and f-measure score which are presented in section 3.1 below.

4.1. Evaluation

The suggested methodology will be evaluated using certain quality measures such as *Precision*, *Recall*, and *F-Measure*.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \ldots. (1)$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \ldots. (2)$$

$$\text{F-Measure} = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad \ldots. (3)$$

4.2. Random Forest

Figure 10 below presents a visual illustration of obtained precisions, recalls, and f-measures scores such as Random Forest has reported a highest precision of 0.91. Similarly, a recall of 0.92 has been noted from this algorithm. The f-measure score of 0.93 has been produced by the Random Forest classifier.

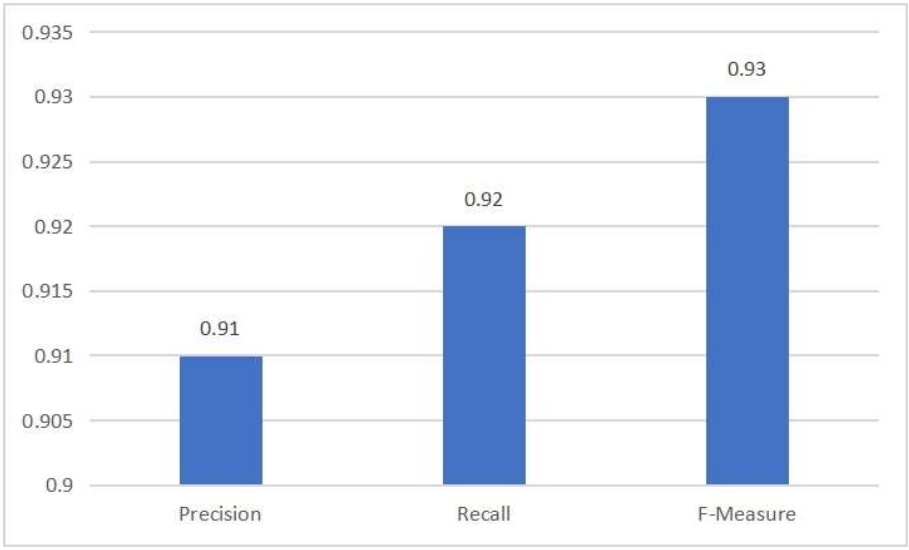

Figure 10: Evaluation of Random Forest based mining approach based on precision, recall and f-measure.

Figure 11 illustrates evaluation score of Random Forest in terms of True Positive, True Negative, False Positive and False Negative rates. In experimentation, a score of 0.291 has been reported as a true positive rate. Likely, a low score of 0.0032 has been noted as a False Positive rate. A score of 0.0036 has been produced by the Random Forest classifier as a True Negative rate. Likewise, in terms of false negative 0.0190 has been observed.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

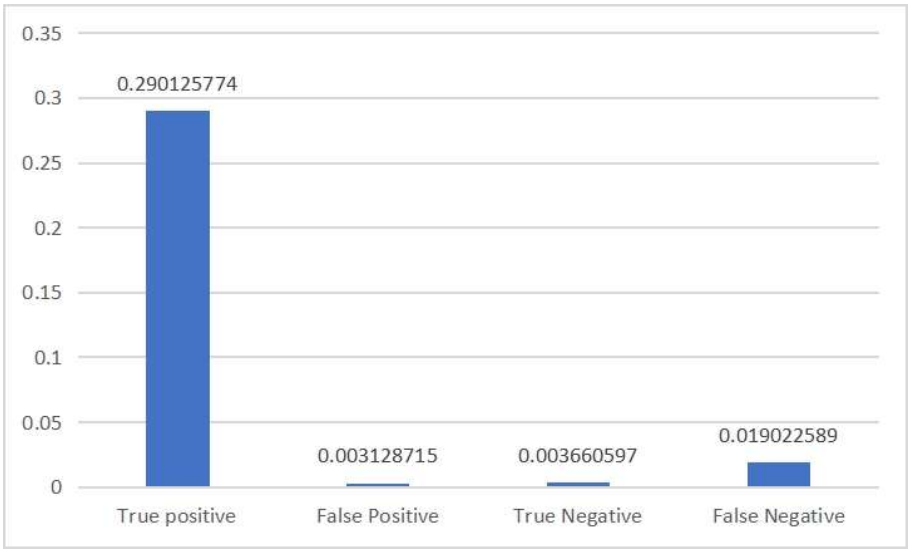

Figure 11: Evaluation of Random Forest based mining approach based on true positive, false positive, true negative and true negative.

4.3. Decision Tree

Figure 12 below present's visualization of obtained precisions, recalls, and f-measures scores such as Decision Tree has reported a precision of 0.87. Similarly, a recall of 0.87 has been noted from this algorithm. The f-measure score of 0.87 has been produced by the Random Forest classifier.

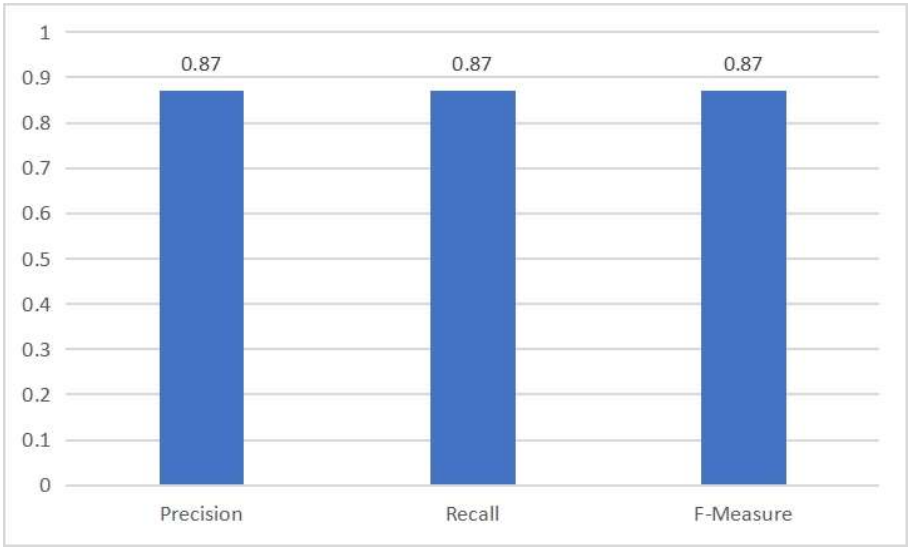

Figure 12: Evaluation of Decision Tree based mining approach based on precision, recall and f-measure.

Figure 13 presents' evaluation score of Random Forest in terms of True Positive, True Negative, and False Positive and False Negative rates. In experimentation, a score of 0.261 has been reported as a true positive rate. Likely, a low score of 0.0189 has been noted as a False Positive rate. A score of 0.0016 has been produced by the Random Forest classifier as a True Negative rate. Likewise, in terms of false negative 0.01878 has been observed.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

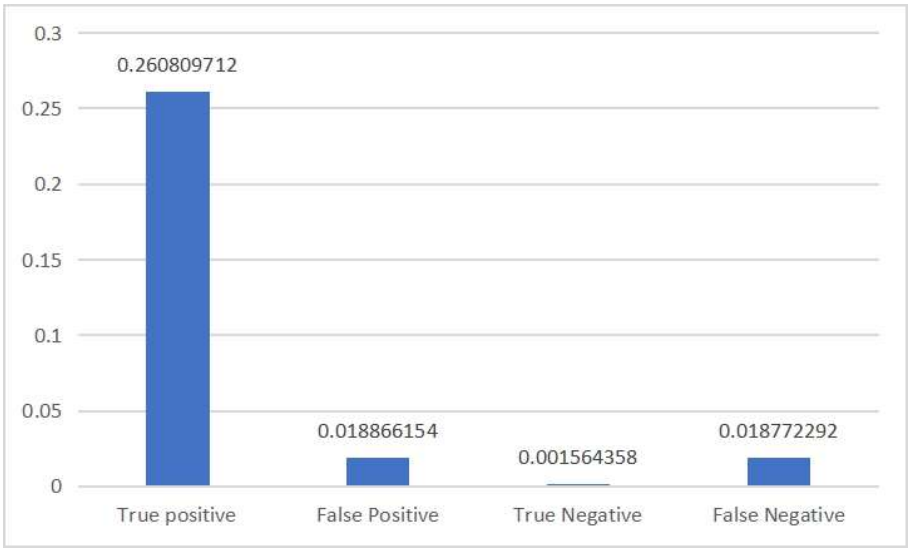

Figure 13: Evaluation of Decision Tree based mining approach based on true positive, false positive, true negative and true negative.

4.4. Linear Regression

Figure 14 below illustrates obtained precisions, recalls, and f-measures scores such as Linear Regression has reported a precision of 0.87. Similarly, a recall of 0.93 has been noted from this algorithm. The f-measure score of 0.90 has been produced by the Random Forest classifier.

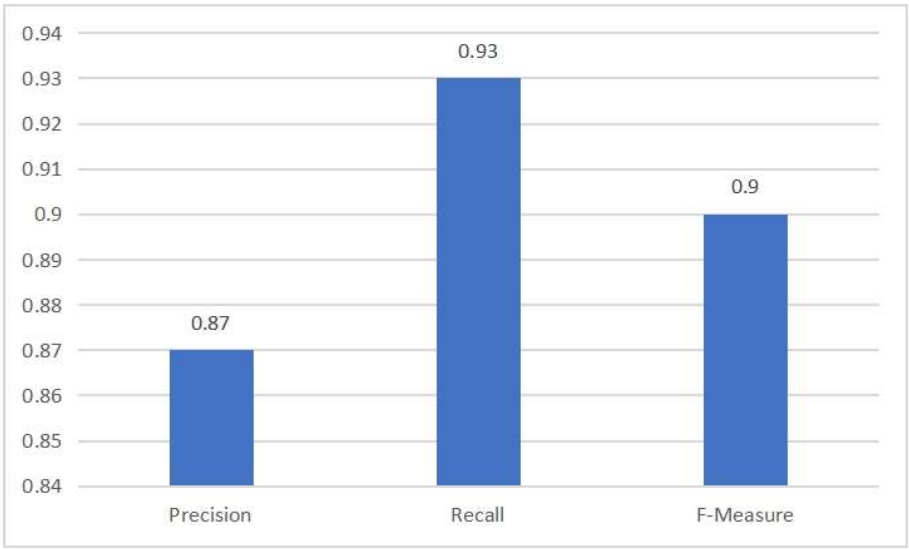

Figure 14: Evaluation of Linear Regression based mining approach based on precision, recall and f-measure.

Figure 15 illustrates evaluation score of Linear Regression in terms of True Positive, True Negative, False Positive and False Negative rates. In experimentation, a score of 0.2797 has been reported as a true positive rate. Likely, a lower score of 0 has been noted

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

as a False Positive rate. A score of 0.0203 has been produced by the Random Forest classifier as a True Negative rate. Likewise, in terms of false negative 0 has been observed.

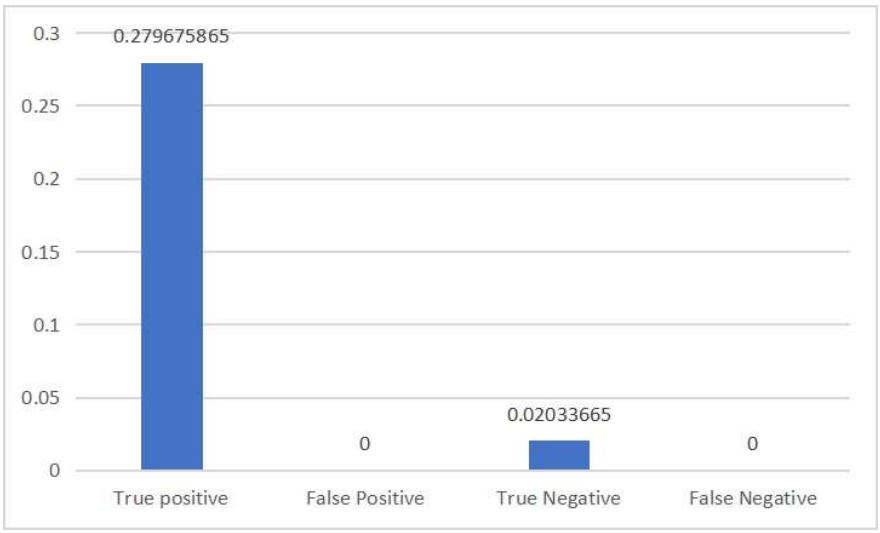

Figure 15: Evaluation of Linear Regression based mining approach based on true positive, false positive, true negative and true negative.

4.5. Comparison of Results

Figure 16 below presents the comparison of the obtained precisions, recalls, and f-measures such as the precision of 0.91 has been obtained from Random Forest, 0.87 from Decision Tree, 0.91 from Linear Regression whereas both reported scores are congruent. Likewise, a recall of 0.92 has been recorded from a random forest whereas the recall of 0.92 from Linear Regression, 0.87 from Decision Tree has been noted. The f-measure scores of 0.93, 0.87 and 0.87 have been recorded from the Random Forest, Decision Tree and Linear Regression, however, there is a tie in between Random Forest and Linear Regression as both have reported congruent scores.

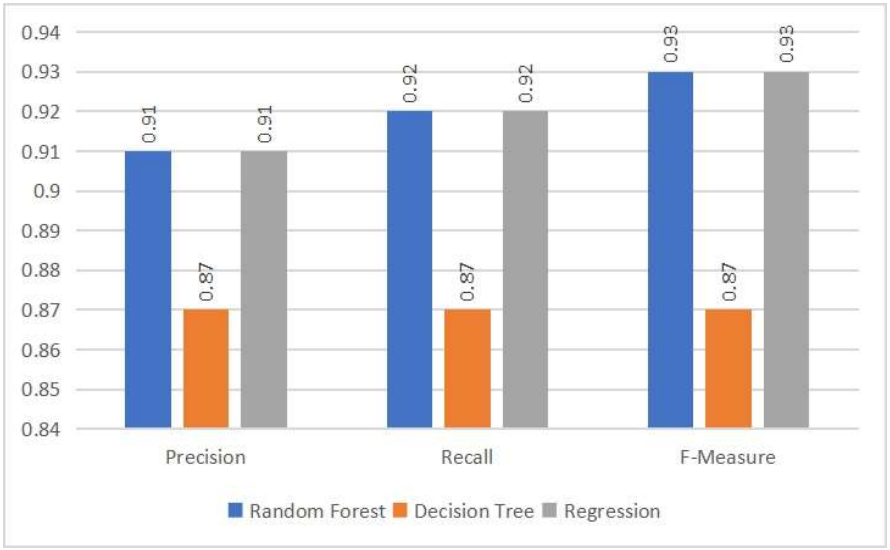

Figure 16: Comparison of employed techniques in terms of precision, recall and f-measure.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

Figure 17 illustrates a comparison of the obtained true positive, true negative, false positive and false positive scores such as a true positive rate of 0.290 has been obtained from Random Forest, 0.26 from Decision Tree, 0.290 from Linear Regression whereas reported scores are congruent in terms of Random Forest and Linear Regression. Likewise, a false positive of 0.003 has been recorded from a random forest whereas the recall of 0.003 from Linear Regression and 0.0018 from Decision Tree has been noted. The true negative scores of 0.003 has been recorded from a random forest whereas the recall of 0.003 from Linear Regression and 0.0015 from Decision Tree has been noted. Furthermore, 0.0190 has been recorded from a random forest and 0.0190 from Linear Regression as a false negative score, however, 0.0018 from Decision Tree has been noted. The discussion in this section delineate that both methodologies such as Random Forest and Linear Regression both can efficiently classify tweets on the basis of their sentiment classes.

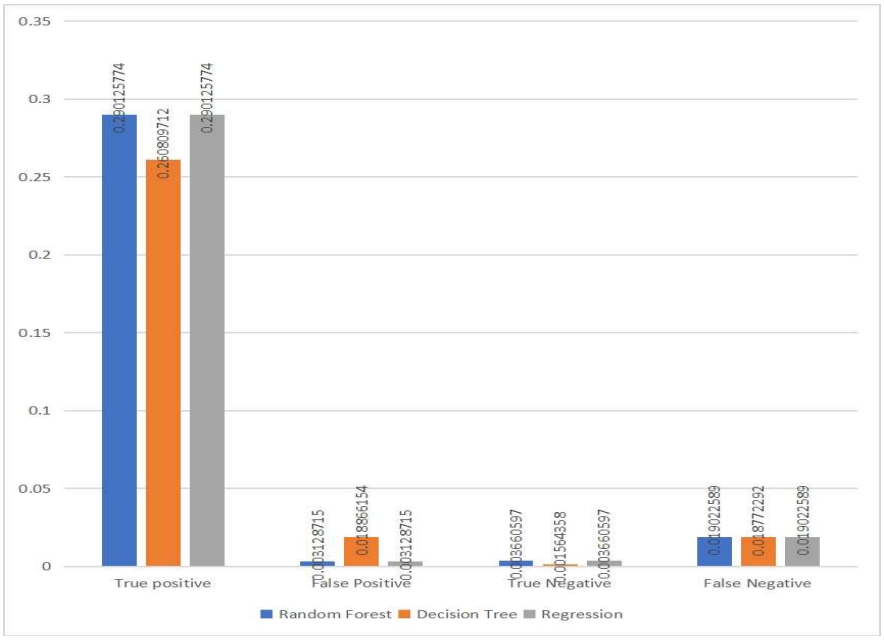

Figure 17: Comparison of employed techniques in terms of true positive, true negative, false positive and false positive.

CONCLUSION

Data mining is an efficient approach which has proven to be effective in analyzing and visualization of a particular dataset. The approach of mining plays a paramount role in making a particular decision over a query. The said approach uses multiple statistical and machine learning based methods to mine the hidden data patterns in comprehensive and diversified datasets. Most of mining techniques make decisions on the basis of artificial intelligence. In other words, artificial intelligence is a backbone of data mining, as it gives the ability of making autonomous decisions using a particular dataset. Artificial Intelligence can be efficiently used in data formatting such as detecting extreme values, outliers, missing values etc. In this study, we addressed the issue of sentiment analysis using AI based mining techniques. To solve this problem, we have used a large corpus of tweets which contains more than 35000 data instances of positive and negative tweets. During the processing of this dataset, we have observed multiple types of punctuations, html tags etc which were significantly removed using AI based mining techniques. We have performed the process of classification using multiple AI based mining

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

techniques such as Random Forest, Decision Tree and Linear Regression classifiers. 66% of the data was used for training however, other was employed for testing purposes which is a supervised approach. Later, all aforementioned approaches were implemented using a powerful tool of Python which has enhanced the accuracy of class predictability. As per the study findings, Random Forest and Linear Regression both can effectively classify tweets as positive and negative respectively.

REFERENCES

Agarwal, A. e. a., 2020. *Sentiment analysis of twitter data.* S. l., workshop on language in social media.

Aiswarya Iyer, R. S., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP),* 5(1), pp. 1-14.

Al-Smadi, M. A.-A. M. J. Y. Q. O., 2018. *Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features.* S. l., s.n.

Ashima Yadav, D. K. V., 2019. Sentiment analysis using deep learning architectures. *Artificial Intelligence Review,* 53(6), pp. 4335-4385.

Bogumił Kamiński, M. J. P. S., 2017. A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research,* 26(1), pp. 135-159.

Deepti Sisodia, D. S. S., 2018. Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science,* Volume 132, pp. 1578-1585.

Feldman, R., 2017. Techniques and applications for sentiment analysis. *Communications of the ACM,* 56(1), pp. 76-88.

Freedman, D. A., 2009. *Statistical Models: Theory and Practice,* S. l.: Cambridge University Press.

Gonçalves, P., 2019. *Comparing and combining sentiment analysis methods.* S. l., ACM conference on online social networks.

Ho, T. K., 2010. *Random Decision Forests.* Montreal, QC, 3rd International Conference on Document Analysis and Recognition.

Kotenko, I. C. A. K. D., 2017. *Evaluation of text classification techniques.* s.l., s.n.

Lin, C. a. Y. H., 2018. *Joint sentiment/topic model for sentiment analysis.* s.l., ACM conference on Information and knowledge management.

Liu, B., 2017. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies,* 5(1), pp. 1-167.

Liu, B., 2020. *Sentiment analysis and opinion mining.* S. l., Synthesis lectures on human language technologies.

Li, X., Liu, B. & Yu, P. S., 2017. *Web data mining: exploring hyperlinks, contents, and usage data.* Berlin, Heidelberg, s.n.

Medhat, W. A. H. a. H. K., 2018. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journa,* 6(1), pp. 4444-4459.

Nti, I. K. A. F. A. a. B. A. W., 2019. Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence from Ghana. *Applied Control Systems,* 12(1), pp. 33-42.

Patil, S. G. A. N. M., 2017. *Convolutional neural networks for text categorization with latent.* s.l., s.n.

Prabowo, R. a. M. T., 2019. Sentiment analysis: A combined approach. *Journal of Informetrics,* 1(191-210), p. 3.

Prastyo, P. H. e. a., 2020. Tweets responding to the Indonesian Government's handling of COVID-19: Sentiment analysis using SVM with normalized poly kernel. *Journal of Information Systems Engineering and Business Intelligence,* 6(2), pp. 112-122.

Sachdeva, K. K. & Sabharwal, A., 2018. *Face recognition using neural network with SURF.* S. l., s.n.

Sajda, P., 2018. Machine learning for detection and diagnosis of disease. *Annual Review of Biomedical Engineering,* Volume 8, pp. 14-57.

Shafi Habibi, M. A. S. A., 2015. Type 2 Diabetes Mellitus Screening and Risk

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.

Factors Using Decision Tree: Results of Data Mining. *Global Journal of Health Science,* 7(5), pp. 304-310.

Shuja Mirza, S. M. M. Z., 2018. Applying Decision Tree for Prognosis of Diabetes Mellitus. *International Journal of Applied Research on Information Technology and Computing,* 9(1), pp. 15-20.

Taboada, M. e. a., 2019. Lexicon-based methods for sentiment analysis. *Computational linguistics,* 37(2), pp. 232-241.

Tatemura, J., 2018. *Virtual reviewers for collaborative exploration of movie reviews.* s.l., s.n.

Tianhua Chen, C. S. P. S. G. A. Q. S., 2018. Effective Diagnosis of Diabetes with a Decision Tree-Initialised Neuro-fuzzy Approach. In: *UK Workshop on Computational Intelligence.* UK: Springer Nature, pp. 227-239.

Vega, L. & Mendez-Vazquez, A., 2018. *Dynamic neural networks for text classification.* s.l., s.n.

Whitelaw, C. N. G. a. S. A., 2019. *Using appraisal groups for sentiment analysis.* s.l., Proceedings of the 14th ACM international conference on Information and knowledge management.

Zadeh, A. e. a., 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint ,* 5(2), pp. 111-121.

Zainuddin, N. & Selamat, A., 2019. *Sentiment analysis using support vector machine.* s.l., s.n.

Corresponding author: Mohammed Musa
ken.mcgarry@sunderland.ac.uk
Faculty of Technology, School of Computer Science, University of Sunderland, United Kingdom.