Digital Innovations
& Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 10. No. 3, September 2022

Article Citation Format
Sarumi, J. A. & Oduroye, A.P. (PhD) (2022):
Towards the Statistical Modelling of Machine Learning
Parameters for Sentiment Analysis of Social Media Messages
Journal of Digital Innovations & Contemporary Research in
Science, Engineering & Technology.
Vol. 10, No. 3. Pp 1-16

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 11th May, 2022
Review Type: Blind
Final Acceptance: 26th July, 2022

Towards the Statistical Modelling of Machine Learning Parameters for Sentiment Analysis of Social Media Messages

Sarumi, J. A[1] (PhD), Oduroye Ayorinde P [2](PhD)
Department of Computer Science
Lagos State University of Science & Technology
Ikorodu, Lagos State, Nigeria.
Caleb University, Imota, Lagos State, Nigeria
E-mail: jerrytechnologies@yahoo.co.uk
Telephone: +2348023408122

ABSTRACT

We present the series of steps involved in the statistical modelling of machine learning parameters for the development of a sentiment analysis framework for social media messages Data collection and dataset creation is presented as the first step towards the creation of the a statistical model using machine learning. The dataset is then divided into three subsets: a training set, a validation set and a test set. The training set is used to train the statistical model, the validation set is used to estimate how well the model is trained and the test set is used to measure the performance of the model. We conclude by presenting the Model Architecture

Keywords: Statistical, Model, Machine Learning, Sentiment Analysis, Social Media

1. INTRODUCTION

Data collection and dataset creation are the first step when creating a statistical model using machine learning. The dataset is commonly divided into three subsets: a training set, a validation set and a test set. The training set is used to train the statistical model, the validation set is used to estimate how well the model is trained and the test set is used to measure the performance of the model. The proposed system consists of four modules - (1) Data pre-processing module: for pre-processing the data (2) Feature representation module: for extracting out features from pre-processed tweets (3) Sentiment classification using base classifiers: in which different base classifiers are used for sentiment analysis and finally (4) Sentiment classification using ensemble classifier.

The details regarding each module is presented below. In the first step, the problem and the objectives for the research is defined. In the second step a literature review is done. The literature study focus on reviewing related work as well as gaining knowledge about the techniques that will be used in the project. In the third step, the experiment setups and configurations will be designed and data will be collected. In the forth step, a prototype tool is developed in order to collect, prepare and analyze data. The analysis is based on mood and sentiment word lists. For the machine learning components in this project the Weka data mining tool is used. In the fifth step, the results are evaluated by measuring the accuracy of performance prediction.

The work in this research is done through five steps, as illustrated in Figure 1

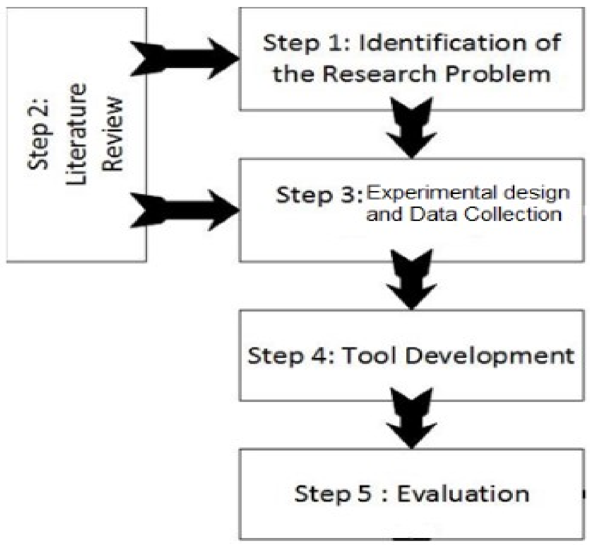

Figure 1: Proposed Methodology

2. METHODOLOGY WORKFLOW

Data Collection

Data collection and dataset creation is the first step when you want to create a statistical model using machine learning. The dataset is commonly divided into three subsets: a training set, a validation set and a test set. The training set is used to train the statistical model, the validation set is used to estimate how well the model is trained and the test set is used to measure the performance of the model.

Tweet Type	Count
Original Tweets	10189
Retweets	6287

Fig 2: Number of original and retweeted Tweets

Data Cleaning

After tweets are gathered from the social network using twitter API based on the query string hash tags, we prepared dataset for sentiment analysis.

i. Collect the tweets that are describing a particular topic from the dataset.
ii. Remove retweet entities, URL removal, markup removal, and hash tags removal. For each given set of tweets, we removed punctuation, numbers, white spaces, and unnecessary symbols

Data Balancing

If the number on instances in classification categories in a dataset are having a huge difference, the dataset is called imbalanced. To counter the issues of imbalanced data, methods such as over-sampling (creating new samples of a certain class) and under-sampling (removing instances of a class) have been proposed. Synthetic Minority Oversampling TEchnique (SMOTE) (Nitesh et al., 2017) is an over-sampling algorithm which provides more instances of the class with lower number of instances in addition to under-sampling of the class with more number of instances. In SMOTE, based on the required number of over-sampling K number of the nearest neighbor to the data point is selected and then after these steps the synthetic sample will be created:

- Take the difference of a data instance to its nearest neighbor,
- Multiply the number by a random value between 0 and 1,
- Add the new data point to the considered feature vector

Data Processing

Data pre-processing module is responsible to decrease the size of the feature set to make it suitable for learning algorithms. This is required because a tweet may contain several features as shown in a sample tweet in Figure 3.1. Following are the steps in the data pre-processing:

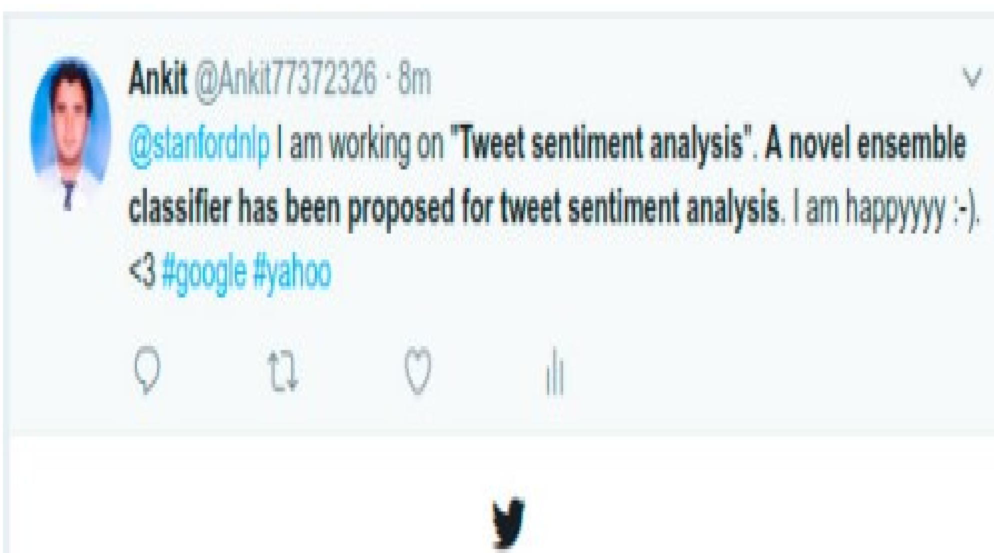

Figure 3: A sample tweet with various features

❖ Retweets, which starts with "RT" are eliminated.

- User names preceded by '@- and external links are eliminated.
- Hashtag '#- (used to point subjects and phrases that are currently in trending topics) is removed from the tweet.
- Emoticons1are replaced by its equivalent meaning because these can serve as a useful feature to detect sentiments.
- "Stemming" is done to reduce each word to its root word.
- Slangs are converted to words with equivalent meaning.
- Stop-words or useless words are removed from the tweet.

Figure 3 illustrates the main steps of the data processing. The starting point is a set of tweets which was extracted via Twitter API. Based on sentiment analysis approach, a sentiment classifier will be built by learning from previously annotated subset of tweets in order to classify the rest of tweets.
The classifier to be built will be able to learn the defined sentiment: "Positive" and "Not Positive". The processing steps are divided into three main steps: Tweets' text filtering, feature extraction, and sentiment classification.

Tweets Text Filtering

As previously mentioned, tweets are informal sentences that have to pass through a filtering stage before it can be processed for the upcoming steps. Filtering is the process of cleaning the tweets text removing all irrelevant text for the sentiment classifier learning step. The following are tweets filtering steps mentioned in the order they were performed

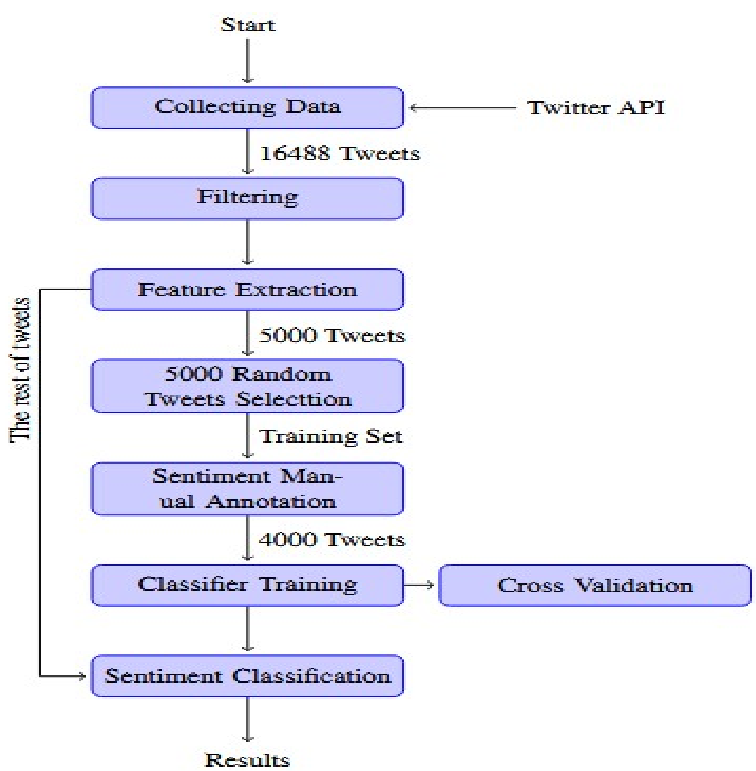

Figure 4: Data Processing Flow

i. All text is switched to lowercase including those words which are completely capitalized. Despite the fact that some users tend to emphasize specific words with capitalization, this was not the general case with the collected tweets. Many names and sentences are found completely capitalized indicating no emphasizing on the meaning. Taking the capitalization into account in such cases might lead to false results. Therefore, all text is changed to lowercase.
ii. All hyperlinks are removed. Tweets mostly contain hyperlinks to other sites and photos which does not contribute to the sentiment of the tweet.
iii. All mentioned usernames (identified by words that start with @) are removed and all hashed words with the # symbol are replaced with the word itself. These specific symbols and markups mentioning usernames or include hashed words that tag a place, name, etc. are so general to con-
tribute to a specific tweet sentiment.
iv. The "RT" text which indicates a retweet is removed.
v. Repeated lettered are filtered. Often, users emphasize words by repeating letters such as: "I am Happyyyy". Alec et al. (2019) suggest to remove out repeated letters leaving only two of them. This also guarantees that words such as "cool" with original double letters are left unaffected.
vi. Common emoticons are replaced with their semantic. Emoticons are often used in social media language to indicate the users' emotions Pak & Paroubek, 2020). The found emoticons are classified as:
vii.

Happy emoticons: ":)", ":-)", ":D", ";)", ";']", etc. which are replaced by "HAPPY FACE".

Sad emoticons: ":-(", ":(", "=(", ";(", ":[", etc. which are replaced by "SAD FACE"

viii. Negations detected in the tweet. Depending on the language, negation appears in different forms. Accordingly, the sentiment of the words appear before and after the negation are changed. For example, "I don't like exams" is changed to "I NOT do NOT Like exams".
ix. All words which do not start with a letter are removed. This eliminates all phone numbers and dates included in the tweet.
x. Extra spaces and punctuation marks are removed.
xi. All stop words and keywords (including the universities' names) are removed based on the language of the tweet

Feature Vectors

A feature vector is the way an object presented in machine learning and pattern recognition. Feature vectors are n-dimensional vectors where each vector represents an object. A numeric representation of the features (variables) will enhance statistical analysis, therefore many machine learning algorithms requires numerical features.

Feature selection

The process of selecting a subset of features that, should be used to construct the model is called feature selection. In machine learning and statistics, the process is also called variable selection. There are various ways to do feature selection. As an example, information gain IG specify the most important features following the formula:

$$p(c|x) - \frac{p(c)p(x|c)}{p(x)} \qquad (1)$$

where:
T is set of training example,
a is the index of a feature
H() function is an entropy (Entropy is a measure of the randomness of a variable and it measures the level of impurity in a group of examples,

Features Selection and Extraction
An important part of the sentiment analysis process is features selection. Features are the sentence properties that are analyzed in an attempt to correlate it to the tweet sentiment (i.e. "Positive" or "Not Positive"). A feature can be the fact that the tweet contains a word, emoticon, a combination of words, etc. The selection of the features is very important as they act as the input for the classifier in the next step. In the features extraction step, they take part in forming the unigrams and bigrams features of the tweets. This leads to a better performance for the classifier (Alec, 2019)

Sentiment Classification
Different classifiers are been presented in the literature for Twitter sentiment analysis. Supervised classifiers are the focus. They require a training set to be prepared forehead. The training set have to be annotated. Pak and Paroubek 2020 did this automatically base on the fact that all tweets in their dataset contains emoticons. They labelled each tweet based on the emoticon sentiment to be either "Positive" or "Negative". For the training set, it got annotated manually by different people based on their own feeling whether the tweet indicate a "Positive" or "Not Positive" sentiment. Automatic labelling was not possible at this stage as the tweets lack a common feature for sentiment labelling. Several classifiers can be used when it comes to the tweets sentiment analysis. Mainly, three common classifiers in the field of machine learning have been used in the literature: Naive Bayes classifier, Support Vector

Machines (SVM), and Maximum Entropy. Naive Bayes and SVM have been compared by Pak and Paroubek 2020 and Alex et al. 2019; Naive Bayes has performed better. Theoretically, Maximum Entropy performs better than Naive Bayes as it handles feature overlap better. However, in practice, Naive Bayes showed better performance on a variety of problems (Alec et al., 2019). Naive Bayes classifier is adapted by this paper's approach. It is a common method for text categorization. It appeared often for solving the problem of determining the category or class of documents that belongs to using word frequencies as the features. In machine learning, Naive Bayes classifier belongs to the family of probabilistic classifiers based on applying Bayes' theorem with the assumption that features are conditionally independence from each other given a specific class

$$P(s|f) - \frac{P(s).P(f|s)}{P(f)} \qquad (2)$$

Equation 1 shows the basic formula of Naive theorem where s is the sentiment class (i.e. "Positive" or "Not Positive") and f is a specific feature. This equation computes the probability of having a tweet with the sentiment s when it contains the feature f . It is calculated based on the probability of having a specific sentiment, probability of the feature existence in all tweets, and the probability of finding the feature in the tweets that belongs to that specific sentiment.

Feature Representation

This module is responsible to extract features from preprocessed tweets. In this paper, Bag-of-Words technique (Han & Kamber, 2016) is used to convert training tweets into numeric representation. Bag-of-Words(BOW) learns a vocabulary of known words from all of the tweets (Yoon et al., 2018). After learning vocabulary, BOW describes the presence of known words within a tweet. For example, consider the following three tweets:

Tweet1: "yesterday is past"
Tweet2: "today is present"
Tweet3: "tomorrow is future".

The vocabulary is {yesterday, is, past, today, present, tomorrow, future}

Now, the above tweets are represented as:
tweet1 vector: [1 1 1 0 0 0 0] tweet2 vector: [0 1 0 1 1 0 0] tweet3 vector: [0 1 0 0 0 1 1]

The parameter values of the BOW are tuned as:
analyzer = "word", ngram range = (1, 2), max features = 4000

3. MACHINE LEARNING

Machine learning is a field of computer science which studies and explores ways of making algorithms find patterns or learn how to do certain tasks. In this thesis machine learning is used to predict the performance of a company. Figure 3.2 shows the workflow for the machine learning process we have used in this thesis.

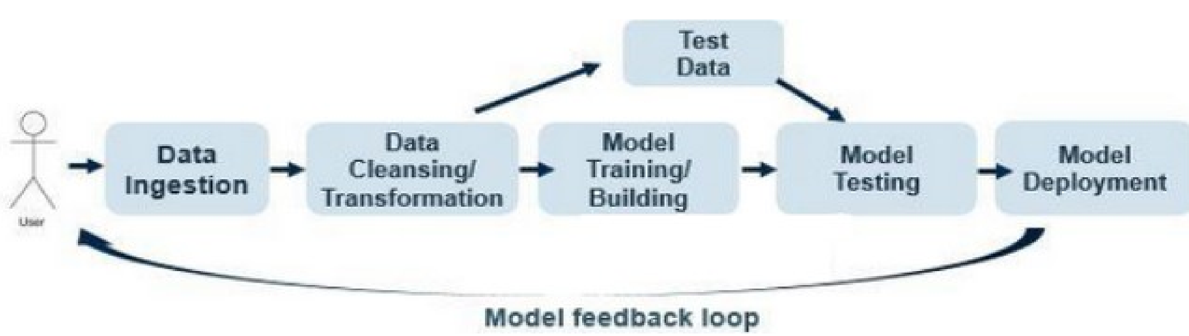

Figure 5: Machine learning workflow (Carol et al., 2015)

A. Supervised Learning: In supervised learning the computer receives a set of inputs and their related outputs from a teacher. The goal is to find a general mapping model from input to output.

B. Unsupervised Learning: In unsupervised learning, the computer find structures in the input data without having any input from a teacher.

C. Reinforcement Learning: In reinforcement learning the computer inter acts with an environment to achieve the goal without any help from a teacher

Classification Algorithms (Base Classifiers)
A classification algorithm task is to pick the right identified categories in data, for the new observations, the classifier estimates categories for new data based on the model parameters that are learned from the training data. Different classification algorithms use different classifier methods and variables and therefore a number of classification algorithms can be applied on the data in order to find the most suitable and efficient algorithm (Karina et al., 2020).

A. Naive Bays is a probabilistic classifier that uses Bayes theory with the assumption that the features are independent (occurrence of one feature. when training model time is important Naive Bays is useful.

$$p(c|x) = \frac{p(c)p(x|c)}{p(x)} \qquad (3)$$

B. AdaBoost (Freund & Schapire, 2016) stands for adaptive boosting and it assumes that finding many weak models are easier than finding one accurate model. Boosting is an approach to create predictions rules with high accuracy using a combination of weak models and rules that have low accuracy in prediction. Boosting generates a sequence of base models and then decides a final estimate of the target variable based on aggregating the estimates of the base models. AdaBoost generates a numbers of weak classifiers and a final estimate of the target variable is chosen based on aggregating the estimates made by the base models. Similar to the random forest algorithm, AdaBoost also have a variable importance estimation but in a different way. In AdaBoost the more informative variables are used more often, and the less informative features are barely used.

C. Cross validation (Sylvain, (2019) creates a training set and a test set by partitioning the original data with the goal to train and evaluate the model. In k-fold cross validation the original data will be divided into k number of subsamples. One subsample is selected as test dataset and the rest (k − 1) number of subsamples are used as training set for the model. The same process will be repeated for k number of times (folds) and each subsample will be used at least once as test set and then the results will be averaged or combined to make the best estimation

4. SENTIMENT ANALYSIS

Sentiment analysis is a method that analyzes how opinions, reactions, impressions, emotions and perspectives are expressed in a language. Its algorithms can extract evaluative information from large text databases and summarize it (Maynard & Funk, (2016). In order to analyze the opinion of people and customers, sentiment analysis appears as the main tool in different contexts.

As an example, sentiment analysis has been used to measure customers satisfaction via statements they comment on a specific product they bought or a service they were delivered. It also appeared in the extent of detecting different opinions regarding political events such as elections. Sentiment analysis methods are; well developed in the domain of blogs and product reviews. Researchers have been working on detecting sentiment in text via presenting different algorithms for detecting semantic orientation. In favour of producing meaningful information from tweets, sentiment analysis used. Different features selection techniques are been investigated, establishing a comparison between different one such as n-grams, part of speech, lexicons, etc. Besides, different classifiers with their learning performance been tested in different contexts (Alec et al., 2019). This paper applies the existing developed approaches in sentiment analysis to microblogging platforms data such as Twitter in order to explore complimentary resources for university evaluation and comparison

Twitter Sentiment Analysis for Evaluating Student Performance

This project suggests that social media content is a vital source for collecting feedback and reactions on the daily events and activities that relates to universities. To prove this hypothesis, a case study is established which evaluates the reactions and feedback from the social media data that is related to the university

Defining Tweets Sentiment

This project approach classifies the sentiment of each tweet to be either “Positive” or “Not Positive”. These, known as two-way sentiment classification (Agarwal et al., 2016). A “Positive” tweet refers to text that indicates a positive statement regarding an event such as a lecture, class or activity that relates to one of the TU9. A “Not Positive” tweet can be either a negative statement regarding an event, or a neutral one, such as an announcement or advertisement regarding an event in the university. Adapting two-way classification can be considered as a limitation. Nevertheless, it is easier to process in the classifier learning step. The same approach was adapted by Alec et al. (2019) who consider that the “Not Positive” tweets are actually “Negative” ones, ignoring the neutral nature of some tweets. Pak and Paroubek proved that adapting a three way classification leads to bad performance which can be avoided by the two way classification (Alec et al., 2019)

Tweets Sentiment Analysis Challenges

Dealing with social media as a source of information - especially microblogging platforms such as Twitter - adds extra difficulties to the sentiment analysis process (Kouloumpis et al., 2016). Tweets are plain text written in an informal manner and its processing face challenges such as:

a) **Length:** Tweets have a limited text length, which is 140 characters. This forces users to start using some common and uncommon abbreviations and phrases. As an example, abbreviations such as OMG5, WTH6, DKDC7, TY8, etc. appears often in twitter
b) **Informality:** Twitter is mostly used as a non-formal communication medium. This leads to many informal statements which probably contain errors such as misspellings, unstructured sentences and slang. Informality may also infer sarcasm, which adds an extra layer of difficulty in guessing the right sentiment of each tweet.
c) **Credibility:** This paper’s approach of gathering the tweets is based on a list of keywords. This does not guarantee the credibility of who and what tweets are generated on Twitter. This leaves the possibility that one anonymous user has generated all the content about a specific university with different usernames, rather than the students.

d) Data availability: collecting the right data is always a challenge, but having enough data is another critical issue. The target data for this study is very specific, which can be problematic for collecting data over a six-month period of time. It is significant to note that more data leads to more trusted results

5. SENTIMENT CLASSIFICATION USING *ENSEMBLE CLASSIFIER*

Base classifiers have been widely used to solve the task of sentiment analysis.

Naive Bayes (NB)

This is a probabilistic classification technique. This classifier performs well when applied to large datasets (Han & Kamber, 2016). NB classifier computes posterior probability by using the formula

$$\text{posterior probability} = \frac{\text{likelihood} \times \text{prior probability}}{\text{evidence}}$$

Equivalently,

$$P(Classi|z) = \frac{p(z|Classi) \times P(Classi)}{p(z)}$$

Where z represents the feature vector and Classi represents the i th class. NB classifier makes an assumption that features are conditionally independent. Smoothing techniques are used to eliminate undesirable effects.

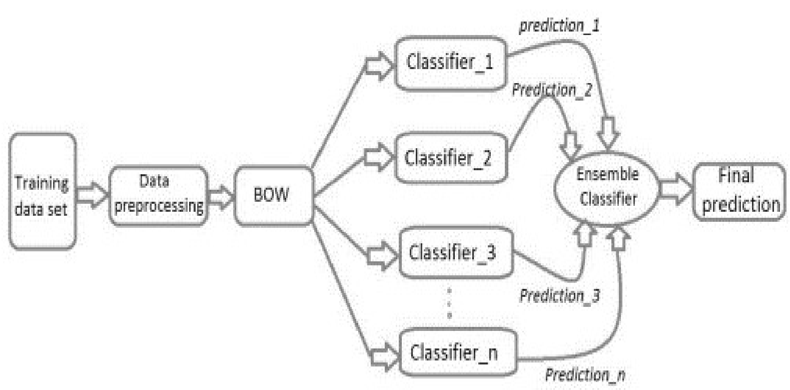

Fig. 6: An Overview of Tweet Sentiment Classification Approach Using Ensemble Classifier

Random Forest (RF)

RF is an ensemble method. Every classifier in the Random Forest is a decision tree classifier. RF classifier builds a set of decision trees from the training dataset (Jianqiang, 2016). After collecting votes from the different decision trees, it decides the final label or class of the test object. The parameter values of the RF classifiers are tuned as: n estimators = 150, max depth = 30. 3.3.3. Support Vector Machine(SVM) This model requires training data to train the model. It is also called a probabilistic classifier (Pang et al., 2017). SVM uses a nonlinear mapping whose aim is to find large margin between different classes.
Although training time of SVM can be slow but it is highly accurate. SVMs attempt to find a decision boundary which maximizes the separation gap between the classes. Unlike Naive Bayes classifier, SVM makes no class conditional independence assumption. SVM yields good result for the task of the Twitter SA problem. The parameter values of the SVM classifiers are tuned as: C = 0.1, kernel = linear.

Logistic Regression (LR)

This is a regression model that is used for classification purpose. LR is generally used to relate a single categorical dependent variable to one or more independent variables (Onan et al., 2016). LR attempts to find a hyper-plane which maximizes the separation gap between the classes. The parameter values of the LR classifiers are tuned as: C = .01, max iter = 100. 3.4. Proposed Ensemble Classifier Ensemble classifier aggregates multiple base classifiers in order to obtain a robust classifier (Prusa et al., 2015). Generally ensemble classifiers have been used to enhance the performance and accuracy of base learning techniques. Figure 2 shows an overview of tweet sentiment analysis approach using the ensemble classifier. Base learners like NB, RF, SVM, and LR are used in ensemble classifier

Algorithm 1: Proposed ensemble algorithm to calculate the Sentiment score of a tweet

1 Function Calculate Sentiment score (Test tweet);
Input : Test tweet
Output: Sentiment score
2 foreach Tweeti in Test tweet do
3 Positive counti = 0
4 Negative counti = 0
5 foreach classifier ci in classifier ensemble do
6 if ci predict Positive then
7 Positive counti += 1;
8 end
9 else
10 Negative counti += 1;
11 end
12 end

$$\text{Probability(Positivei)} = \frac{\text{Positive count}}{\text{Positive counti + Negative counti}}$$

$$\text{Probability(Negativei)} = \frac{\text{Negative counti}}{\text{Positive counti + Negative count}}$$

13 end
14 foreach classifier ci in classifier ensemble do

$$Weight_{c_i} = \frac{acc_{c_i}}{\sum_{j=1}^{n} acc_{c_j}}$$

// Where accci is the accuracy of i th classifier, j denotes the no. of learning classifiers in the ensemble classifier and acccj represents to the accuracy of j th learning classifier.

15 end
16 foreach Tweeti in Test tweet do
17 Positive scorei = 0
18 Negative scorei = 0
19 foreach classifier ci in classifier ensemble do
20 if ci predict Positive then
21 Positive scorei += Weightci * Probability(Positivei);
22 end

```
else
Negative scorei += Weightci * Probability(Negativei);
end
end
return Positive scorei, Negative scorei
end
```

Algorithm 1 calculates sentiment score of the tweet. The system was trained using the training data. The T est tweet is a set of tweets that was used to test the system. Each base classifier in ensemble classifier determines the sentiment (Positive/Negative) of each tweet in T est tweet. In addition, the classification report of each base classifier was calculated on the testing data (T est tweet). The next step is to calculate the probability of each tweet being positive and negative. After assigning this probability, we assign the weight to each classifier in the ensemble technique based on the accuracy of each classifier. Finally, the algorithm calculates the positive and negative score of the tweet based on the prediction of each classifier.

Algorithm 2 predicts the sentiment of the tweet. The inputs to this algorithm are the positive score and negative score of the tweet. If the positive score of the tweet is more than its negative score, then the sentiment of that tweet is taken as positive. And, if the negative score of the tweet is more than positive score then the sentiment of that tweet is taken as negative. Finally, If the positive score and the negative score of a tweet are equal then the system calculates the cosine similarity of that tweet with all other tweets in the testing data and identifies the most similar tweet. Then it calculates the positive and negative score of the identified tweet. Now if positive score is more than negative score then tweet is positive otherwise it is taken as negative.

Algorithm 2: Proposed ensemble algorithm to predict the sentiment of a tweet

```
function SentimentPredictor (Tweeti, Positive scorei, Negative scorei);
Input : Tweeti, Positive scorei, Negative scorei
Output: S entiment
if Positive scorei > Negative scorei then
Sentiment = “Positive”;
else
if Negative scorei > Positive scorei then
Sentiment = “Negative”;
else
Calculate cosine similarity of Tweeti with all other tweets in test data using distance calculation formula.
Find the most similar tweet of Tweeti in Test tweet, say Tweetj.
calculate Positive scorej and Negative scorej of Tweetj using Algorithm 1.
if Positive scorej >= Negative scorej then
Sentiment = “Positive”;
else
Sentiment = “Negative”;
end
end
end
Return Sentiment
```

Distance calculation "Cosine similarity" measures the similarity of a pair of tweets. Cosine similarity can be computed by using this formula:

$$\cos(\text{tweet1}, \text{tweet2}) = \frac{\text{tweet1} \cdot \text{tweet2}}{||\text{tweet1}|| \cdot ||\text{tweet2}||}$$

Where tweet1 and tweet2 represent vectors and output value 1 represents high similarity

5.1 Architecture

Figure 7 shows the general framework of our proposed approach. The different components involved in this framework are explained in the subsections.

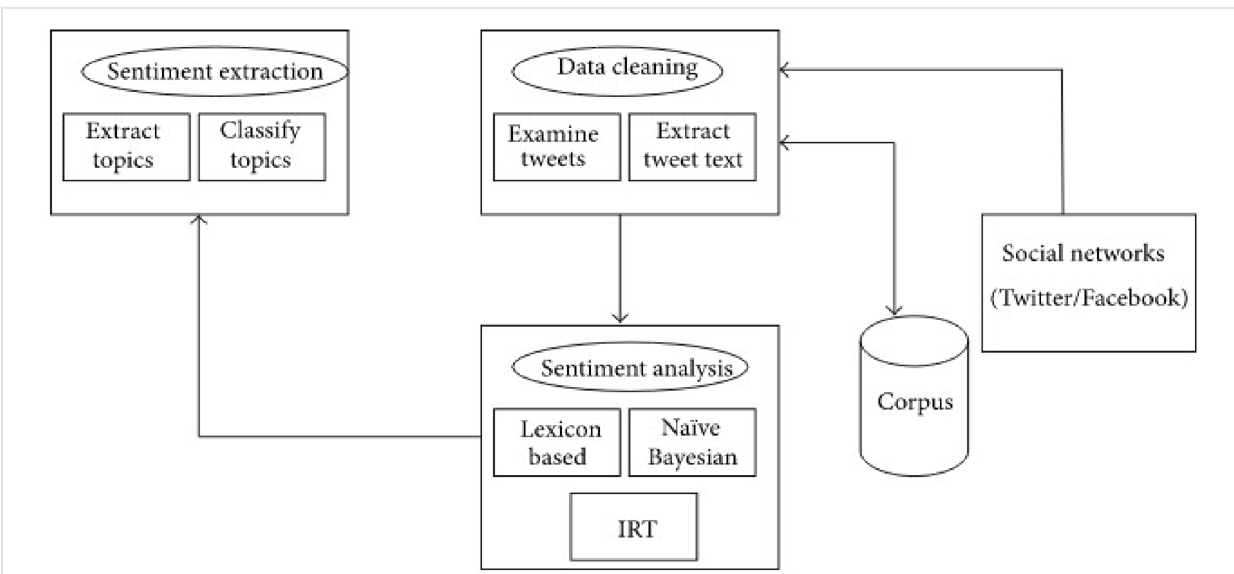

Figure 7: Model Architecture

6. CONCLUSION & Future WORKS

In this paper we provided insights into the series and sequence of steps involved in the statistical modelling of machine learning parameters for the development of a sentiment analysis framework for social media messages. Future works will seek to implement the model

BIBLIOGRAPHY

1. Reza Zafarani, Mohammad Ali Abbasi, Huan Liu (2019). Social Media Mining Cambridge University
2. Buchem, I., Ebner, M., Schön, S., Appelt, R., & Kaiser, S. (2018). Blogging und Microblogging. Lehrbuch für Lernen und Lehren mit Technologien: 2. Auflage (2013), 239.
3. Gapski, H., & Gräßer, L. (2017). Medienkompetenz im web 2.0–lebensqualität als zielperspektive. Praxis Web, 2, 11-34.
4. Hisserich, J., & Primsch, J. (2020). Wissensmanagement in 140 Zeichen: Twitter in der Hochschullehre. Community of Knowledge.
5. Maynard, D., & Funk, A. (2016). Automatic detection of political opinions in tweets. In Extended Semantic Web Conference (pp. 88-99). Springer, Berlin, Heidelberg.
6. Go, A., Bhayani, R., & Huang, L. (2019). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.
7. Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. Computational Intelligence, 31(1), 132-164.
8. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2016). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
9. Kouloumpis, E., Wilson, T., & Moore, J. (2016). Twitter sentiment analysis: The good the bad and the omg!. In Fifth International AAAI conference on weblogs and social media.
10. Pak, A., & Paroubek, P. (2020). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).
11. Davidov, D., Tsur, O., & Rappoport, A. (2020). Enhanced sentiment learning using twitter hashtags and smileys. In Coling 2010: Posters (pp. 241-249).
12. Kaplan, A. M., & Haenlein, M. (2020). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), 59-68.
13. Subhabrata Mukherjee. (2017). Sentiment analysis. Indian Institute of Technology, Bombay Department of Computer Science and Engineering.
14. Hagel, J. I. I. I., Brown, J. S., & Davison, L. (2020). The best way to measure company performance. Harvard Business Review, 4(6), 56-60.
15. Needles, B. E., Powers, M., & Crosson, S. V. (2018). Principles of accounting. Cengage Learning.
16. Granger, C. W. (2019). Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, 424-438.
17. Michael T. Lash and Kang Zhao (2016). Early Predictions of Movie Success: the Who, What, and When of Profitability Artificial Intelligence (cs.AI); Social and Information Networks (cs.SI).
18. Sahar Nassirpour, Parnian Zargham, Reza Nasiri Mahalati (2017). Electronic Devices Sales Prediction Using Social Media Sentiment Analysis Project report Stanford university.
19. Richard Socher Jeffrey Pennington, Eric H. Huang Andrew, Y. Ng Christopher, D. Manning (2016). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions Proceedings of the Conference on Empirical Methods in Natural Language Processing Pages 151-161
20. Tristan Fletcher, Fabian Redpath and Joe DAlessandro (2019) Machine Learning in FX Carry Basket Prediction Proceedings of the International Conference of Financial Engineering, vol. 2, page 1371-1375.
21. Tim Loughran and Bill McDonald (2016) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks The Journal of Finance, Vol. LXVI, NO. 1

22. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining.In LREc (Vol. 10, No. 2010, pp. 1320-1326).
23. Douglas M. McNair, Maurice Lorr, and Leo F. Droppleman (2021). Manual for the Profile of Mood States San Diego, CA: Educational and Industrial Testing Service
24. Carol McDonald. (2015). Parallel and Iterative Processing for Machine Learning Recommendations with Spark https://www.mapr.com/blog/ parallel-and-iterative-processing-machine-learning-recommendations-spark
25. Karina Gibert, Miquel Snchez-Marr, Vctor Codina. (2020). Principles of Accounting. International Environmental Modelling and Software Society (iEMSs).
26. Rokach, L., & Maimon, O. Z. (2017). Data mining with decision trees: theory and applications (Vol. 69). World scientific.
27. Freund, Y., & Schapire, R. E. (2016). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).
28. Sylvain Arlot, (2019). A survey of cross-validation procedures for model selection. Journal of Machine Learning Research , p1089-1105.
29. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. (2017). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, page 321357.
30. Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2016). Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First workshop on Unsupervised Learning in NLP (pp. 53-63).
31. Medhat, W., Hassan, A., & Korashy, H. (2019). Sentiment analysis algorithms and applications:
32. A survey. Ain Shams engineering journal, 5(4), 1093-1113.
33. Davidov, D., Tsur, O., & Rappoport, A. (2020). Enhanced sentiment learning using twitter hashtags and smileys. In Coling 2010: Posters (pp. 241-249).
34. Read, J. (2015). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL student research workshop (pp. 43-48).
35. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2016). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. HP Laboratories, Technical Report HPL-2011, 89.
36. Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Systems with Applications, 62, 1-16.
37. Yoon, V. Y., Hostler, R. E., Guo, Z., & Guimaraes, T. (2018). Assessing the moderating effect of consumer product knowledge and online shopping experience on using recommendation agents for customer loyalty. Decision Support Systems, 55(4), 883-893.
38. Lin, J., & Kolcz, A. (2017). Large-scale machine learning at twitter. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 793-804).
39. Rodriguez-Penagos, C., Atserias, J., Codina-Filba, J., García-Narbona, D., Grivolla, J., Lambert, P., & Saurí, R. (2018). Fbm: Combining lexicon-based ml and heuristics for social media polarities. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2018) (pp. 483-489).
40. Hassan, A., Abbasi, A., & Zeng, D. (2018). Twitter sentiment analysis: A bootstrap ensemble framework. In 2018 international conference on social computing (pp. 357-364). IEEE.

41. Clark, S., & Wicentwoski, R. (2018). Swatcs: Combining simple classifiers with estimated accuracy. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2018) (pp. 425-429).
42. Jianqiang, Z. (2016). Combing semantic and prior polarity features for boosting twitter sentiment analysis using ensemble learning. In 2016 IEEE first international conference on data science in cyberspace (DSC) (pp. 709-714). IEEE.
43. Pang, B., Lee, L., & Vaithyanathan, S. (2017). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.
44. Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Systems with Applications, 62, 1-16.
45. Prusa, J., Khoshgoftaar, T. M., & Dittman, D. J. (2015). Using ensemble learners to improve classifier performance on tweet sentiment data. In 2015 IEEE International Conference on Information Reuse and Integration (pp. 252-257). IEEE.
46. Coletta, L. F., da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. (2019). Combining classification and clustering for tweet sentiment analysis. In 2014 Brazilian conference on intelligent systems (pp. 210-215). IEEE.