Sentiment Analysis with ML Techniques: Handling Imbalanced Dataset

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF DEGREE
OF

MASTER OF TECHNOLOGY IN
SOFTWARE ENGINEERING

Submitted By:

VASUNDHARA RAJ

2K18/SWE/17

Under the supervision of

MR. RAHUL

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

JUNE, 2020

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)
Bawana Road, Delhi-110042

DECLARATION

I, Vasundhara Raj, Roll No. 2K18/SWE/17 student of M.Tech (Software Engineering), hereby declare that the Project Dissertation titled "**Sentimeent Analysis with ML Techniques: Handling Imbalanced Dataset**" which is submitted by me to the Department of Computer Science & Engineering , Delhi Technological University, Delhi Report of the Major II which is being submitted to Delhi Technological University, Delhi, in partial fulfillment for the requirement of the award of degree of Master of Technology for the requirements of the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: DTU, Delhi

Vasundhara Raj

Date: 25-07-2020

(2K18/SWE/17)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**Sentimeent Analysis with ML Techniques: Handling Imbalanced Dataset**" which is submitted by Vasundhara Raj, Roll No. 2K18/SWE/17, Department of computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment for the requirement of the award of degree of Master of Technology (Software Engineering) is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 25-7-2020

Mr. Rahul
SUPERVISOR
Assistant Professor
Department of Computer Science and Engineering
Delhi Technological University.

ABSTRACT

Sentiment Analysis is a process of analyzing and categorizing the emotion or sentiment over any given review or text piece in order to know what the reviewer wants to express in the form of positive, negative or neutral. Today, people are highly interested in buying things online from any e-commerce site or they search for a product review in order to know the quality and one's perception toward that product before buying. Same goes for one wanting to download an App and would definitely view the reviews laid on those review section to know about that App. They would go with the app with highest rating or downloads or one with the good reviews depending on the person's interest. The product/App provider also gets to know about the user's opinion over a product. This can help the company to improve its marketing strategy and quality of product in their favor. Sentiment analysis uses various semantic approaches like on these online reviews to extract as much feature it can and categorize the type of opinion. Some techniques also help in rating the product value based on user's opinion. This project deals with four different ML techniques; Naïve Bayes, Decision Tree, Random Forest and AdaBoost, training the models for classification of sentiment. The main focus of this project work is to handle the imbalanced datasets. In imbalanced dataset, the classes are of skewed size; having classes in majority and minority sizes. These imbalanced dataset affects the performance of the model and the model would behave biased. To improve the performance of the classification model various Sampling Techniques (Random Under Sampling, SMOTE, SMOTEENN, SMOTEToken) can be used. This project handles the same and performance result of model before and after applying sampling technique is depicted. The metric used for comparison between sampling techniques and classification techniques are Precision, Recall and Accuracy. SMOTEENN can been seen outperforming other three techniques. The comparison results have been shown through tables and graphs in section 4.2.

ACKNOWLEDGEMENT

I am most thankful to my family for constantly encouraging me and giving me unconditional support while pursuing this research.

I am extremely grateful to **Mr. Rahul,** Asst. Professor, Department of Computer Science Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to her for the extensive support and encouragement she provided.

I also convey my heartfelt gratitude to all the research scholars of the web Research Group at Delhi Technological University, for their valuable suggestions and helpful discussions throughout the course of this research work.

Vasundhara Raj
Roll No.: 2K18/SWE/17

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

SA	Sentiment Analysis
ML	Machine Learning
AI	Artificial Intelligence
POS	Part-of-speech
NB	Naive Bayesian
NLP	Natural Language Processing
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative
RF	Random Forest
SMOTE	Synthetic Minority Oversampling Technique
DT	Decision Tree
AdaBoost	Adaptive Boosting
ENN	Edited Nearest Neighbor

CHAPTER 1 INTRODUCTION

1.1. RESEARCH OBJECTIVES

Sentiment analysis uses various semantic approaches like on these online reviews to extract as much feature it can and categorize the type of opinion [1]. Some techniques also help in rating the product value based on user's opinion. This project deals with four different ML techniques; Naïve Bayes, Decision Tree, Random Forest and AdaBoost, training the models for classification of sentiment. The main focus of this project work is to handle the imbalanced datasets. In imbalanced dataset, the classes are of skewed size; having classes in majority and minority sizes. These imbalanced dataset affects the performance of the model and the model would behave biased. To improve the performance of the classification model various Sampling Techniques (Random Under Sampling, SMOTE, SMOTEENN, SMOTEToken) can be used. This project handles the same and performance result of model before and after applying sampling technique is depicted. The metric used for comparison between sampling techniques and classification techniques are Precision, Recall and Accuracy.

1.2. ORGANIZATION OF THESIS

The project report has been divided into five chapters. Each chapter deals with one component related to this thesis. Chapter 1 being introduction to this thesis, gives us the brief introduction about the project, thereafter chapter 2 tells about the Background Knowledge of the Sentiment Analysis and Machine Learning Techniques along with a literature survey dealing with sentiment classification using ML techniques as well as handling imbalanced dataset using various Sampling Techniques. Chapter 3 focuses on explaining the imbalanced dataset, why it is a problem and how to handle it. Various sampling techniques used to do this task. This project focuses on four sampling technique; Under-sampling by using Random Under-sampling, Over Sampling using SMOTE techniques and by combining both under-sampling and over-sampling, SMOTEENN and SMOTETomek is considered. Further, Chapter 4 shows various data visualization and result tables and graphs. Conclusion and Future work are discussed in chapter 5.Further Appendix contains list of publication details.

CHAPTER 2 BACKGROUND KNOWLEDGE

The chapter explains the topic Sentiment Analysis and the work done so far in this field of handling dataset imbalance.

2.1. SENTIMENT ANALYSIS

Today with the approach of ML procedures, AI and the expanded mechanization in each division from the IOT based smart gadgets to modern robots, the need is being acknowledged in each other part which has not recovered the true abilities of this innovation. The development of these ideas can be acknowledged from the way that robots were prior just characterized by a mechanical arm and now it has moved to instruct pendant strategies to now even programmable robots. One of the key zones of NLP, a subset of AI is SA, the capacity to comprehend passionate tones in discourse and print. It is a territory that is the concentration for various diverse utilitarian applications.

2.1.1. Why there is a need of SA?

Let's understand it by taking an application of customer service. Success of a company and its service growth directly depends on its client. If the client enjoys the product, then its success of the company else it needs to improvise it by making some changes in it. So in order to know whether customer like your product or not you need to analyse them. One of the attributes is to analyse the customers by their sentiment. That's how the sentiment comes into the picture. So, a procedure of computationally recognizing and categorizing sentiments from a text, and conclude whether the outlook of the writer for the topic or the product is +ve, -ve or neutral. As a user, one does not perform a SA but one does look for response like before purchasing a product one looks into what other customer have to say about that product whether it is decent or bad. And that user analyses it manually. Now, while considering it at a company level, that how will a company analyse what a customer thinks about their service. A company have customer in abundance and so their reviews. This is where a company needs to carry out SA to know if the product or service is actually working in the

market or not. Similarly goes for other applications like Companies running hotels can analyse its service by left out online reviews by customer detailing how was their stay or service at the hotel. Twitter comments section contains a lot of opinions and those can be used to analyse over an issue or topic categorizing polarity and predicting the overall sentiment. Sentiment Analysis can also be used in marketing businesses in call centres too. A customer service bot can be improvised to handle various sentiment of the customer. As a customer can be frustrated or happy by the service or using any slang words or those words which can be treated as both negative and positive in different context, for example: “Not Bad!!”.

2.1.2. Fields that are focused on for Sentiment Analysis

Since there is abundance of research amplifying day by day, so, previous work information would ease some hustle for the researchers as well as students interested in this field of SA and handling imbalanced dataset. The literature survey provided in this work would give a big picture of different areas where Sentiment Analysis is being used and can be improved, along with what different solutions for imbalanced dataset techniques and classification techniques are taken into account by different authors and also some future work aspects and scope in customer service or opinion analysis sector. Some of the application sectors that are being focused are:

- Fraud Detection / Spam Detection
- Twitter dataset that contains abundance of opinion and that is useful is analysed their sentiment. It would help to know the polarity of users regarding a trending topic or posts or what is interesting user so much that could be made trending.
- Facebook comments
- Online Hotel reviews from prominent sites like Trip Advisor, etc.
- Product Reviews from Amazon, Flipkart, Yelp and other e-commerce sites.
- Movie Reviews from Rotten Tomatoes, IMDB, etc. that take into account the review of a movie including its rating, plot story opinion, that would be beneficial for other users to let them know whether they are spending their time on worth their interest or not. Also it would help the production team to know their users taste better and improvise as required.
- News Articles which contains opinions or reviews of a user over a trending topic or news which would highlight the perspective of a particular user or millions of users.

- Healthcare Reviews of some health centres that would describe the facilities provided by the centres and doctors practice reviews, which would help other users to know well about that centre as well as suggest the authority to upgrade their system according to user wellness.

Not just in marketing or business field, social media analytics can likewise have anticipated and clarify the feelings of concerned ideological groups during elections, which has prodded various non-brand associations to explore how SA can be utilized to foresee results and guide out the enthusiastic scene of individuals, voters and so forth. Furthermore, organizations are taking a gander at ways that SA can be utilized outside of their marketing and PR departments. SA essentially glances progressively famous later on.

2.1.3. Sentiment Analysis process

Sentiment Analysis is a process to determine the person's opinion or attitude over any product, service or organization. We will discuss each step in brief as depicted in Fig1.

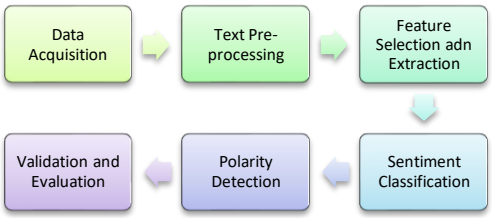


Fig 1. Sentiment Analysis Process Model

1. **Data Acquisition**: The first step of Sentiment Analysis is to acquire or collect data on which sentiment approaches has to be performed. These datasets can be any online review on any product from different sites or any opinion. Without data collection model cannot step further. On these datasets analysis and classification of text can be performed.

2. **Text Pre-processing**: After data acquisition, data needs to be pre-processed which includes removal of unwanted noise, emoticons used, duplicate words, stemming,

different types of URLs, etc. to make data noise-free and extract best features.

3. **Feature Selection and Extraction:** This step is very important in order to get best out of nothing. Filtered data is selected and good extraction technique is used to extract feature for the precision of the model. Different extraction technique used in sentiment analysis is Total Weighted Score Computing Method, Neutral/Polar/Irrelevant Classification Model, Weighing and Aggregation Scheme, Intrinsic and Extrinsic Domain Relevance Approach, etc.

4. **Sentiment Classification**: Now different sentiment analysis techniques are used to classify the extracted feature into various sentiments.

5. **Polarity Detection**: After the sentiment classification step, polarity detection is done to determine the polarity of statement, whether it is positive, negative or neutral.

6. **Validation and Evaluation**: Finally, validation and evaluation of the obtained result is done in order to testify the precision of the result.

2.1.4. Sentiment Analysis Technique

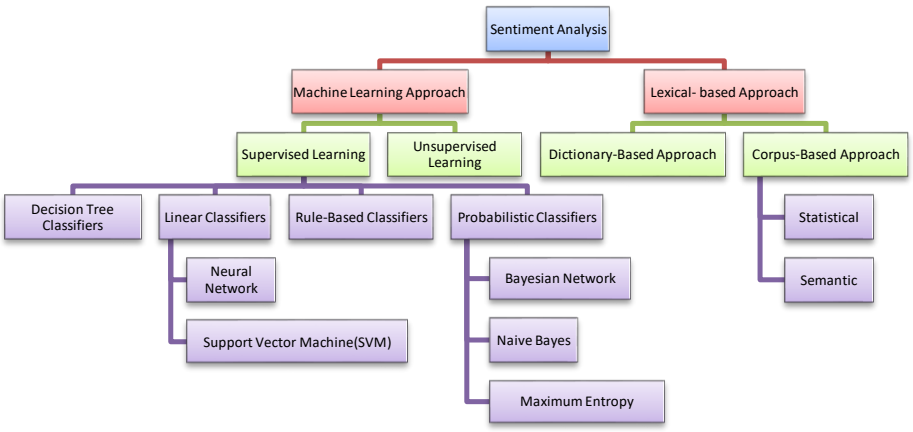


Fig 2. ML Classification Models

Sentiment Analysis uses different techniques for the classification of the extracted feature. These are classified into Machine Learning Approach and Lexical Based Approach. Machine Learning approach includes various other approaches under

Unsupervised and Supervised learning with Support Vector Machine, Naïve Bayes, Decision Tree Classifier and Rule Based Classifier. These algorithms work on sample datasets to determine prognosis or decision. The learning algorithms differ in their approach and depend on input and output data.

2.2. RELATED WORK

T. Lu et.al. worked on intrusion detection system that is useful in security of network. This paper worked on SMOTE and ENN on Random Forest Classifier by selecting 25 features. This method achieved higher precision, recall and F1- value [2].

I. Benchaji et.al. proposed a testing technique dependent on the K-means clustering and the genetic algorithm. We utilized K-means algorithm to bunch and gathering the minority kind of test, and in each group we utilize the genetic algo to pick up the new examples and build a precise extortion location classifier [3].

Y. Pang et.al. took imbalanced network traffic dataset that would tell about the malware detection occurrence. This paper used synthesized approached combining with other random oversampling. It compares this result with ten other resampling techniques and showed efficient result [4].

F. Yin et. al. used the combination of user complaints and the parameters of the network to make a fault prediction model in order to increase revenue of the service providers. It used K- cluster centroid with SMOTE to get K-SMOTE approach which delivered efficient result on decision tree model [5].

S. Zhang used several SVM algorithms for imbalanced dataset classification [6]. X. Li et. al. worked on oversampling technique on minority class in order to increase it and achieve good recognition rate. P. Shukla et.al. in her work dealt with the imbalanced dataset by balancing it by using K- Means and then using SVM for classification, showing better results than working on imbalanced dataset [7].

Other many work has been done in this field in order to handle the imbalanced dataset. Many algorithms have been introduced and used in past works [8] [9] [10] [11] [12] [13].

2.3. MACHINE LEARNING TECHNIQUES USED

Let’s discuss the four machine learning techniques that will be used to analyse the performance of the SA as well as sampling techniques to improve the imbalanced dataset.

2.3.1. Naive Bayes

This technique is simple group of probabilistic algorithm which assigns the probability for a given word or text whether it should be considered positive or negative. The Naive Bayes has three type of models: Gaussian, Multinomial and Bernoulli. Multinomial is used when a word's frequency/ discrete count in a whole document is to be known. Bernoulli NB technique is used where dataset’s feature values are in binary nature. In Gaussian Naive Bayes, persistent qualities related with each feature are thought to be conveyed by a Gaussian distribution. A Gaussian distribution is also known as Normal circulation. At the point when plotted, it gives a bell shaped formed bend which is symmetric about the mean of the values of the features.

2.3.2. Decision Tree

Decision Trees are a non-parametric supervised learning strategy utilized for both classification and regression tasks. The objective is to make a model that predicts the estimation of an objective variable by taking in basic decision rules deduced from the information features.

2.3.3. Random Forest

Random forest is a supervised learning calculation. The "forest" it constructs, is a group of choice trees, generally prepared with the "bagging" technique. The general thought of the bagging technique is that a mix of learning models expands the general outcome. Random forest forms different choice trees and combines them to get an increasingly precise and stable expectation.

One major preferred position of random forest is that it tends to be utilized for both classification and regression issues, which structure most of current ML frameworks. How about we see random forest in classification, since classification is some of the time

considered the building block of ML. Underneath you can perceive how a random forest would look like with two trees:

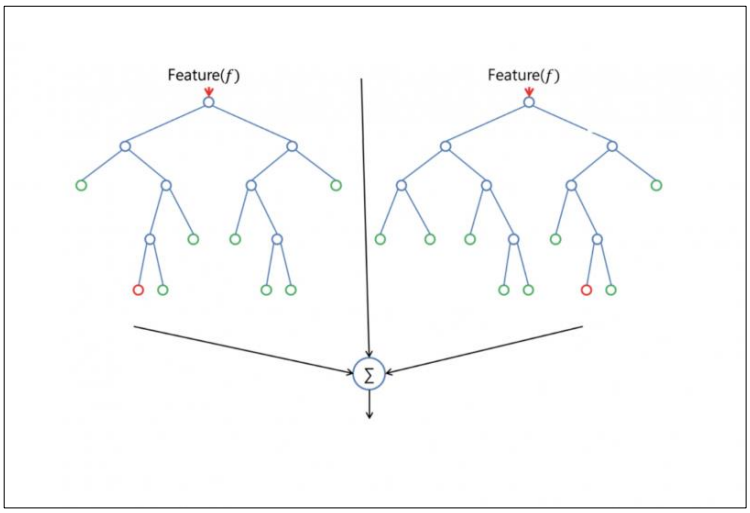


Fig3. Random Forest

Random forest has about the equivalent hyper-boundaries as a choice tree or a bagging classifier. Luckily, there's no compelling reason to consolidate a choice tree with a bagging classifier since you can without much of a stretch utilize the classifier-class of random forest. With random forest, you can likewise manage regression assignments by utilizing the calculation's regressor. Random forest adds extra randomness to the model, while developing the trees. Rather than scanning for the most significant element while parting a hub, it looks for the best component among a random subset of highlights. This outcome in a wide decent variety that by and large outcomes in a superior model. In this way, in random forest, just a random subset of the highlights is thought about by the calculation for parting a hub. You can even make trees increasingly random by moreover utilizing random limits for each component instead of scanning for the most ideal edges (like a typical choice tree does).

2.3.4. AdaBoost

It joins various classifiers to build the accuracy of classifiers. AdaBoost is an iterative troupe strategy. AdaBoost classifier constructs a solid classifier by consolidating different ineffectively performing classifiers with the goal that you will get high accuracy solid classifier. The essential idea driving Adaboost is to set the weights of classifiers

and training the information test in every emphasis with the end goal that it guarantees the precise forecasts of uncommon perceptions. Any machine learning calculation can be utilized as base classifier in the event that it acknowledges weights on the training set. Adaboost should meet two conditions:

1. The classifier ought to be prepared intuitively on different gauged training models.
2. In every cycle, it attempts to give a great fit to these models by limiting training blunder.

It works in the accompanying advances:

- Initially, Adaboost chooses a training subset haphazardly.
- It iteratively prepares the AdaBoost machine learning model by choosing the training set dependent on the exact forecast of the last training.
- It doles out the higher weight to wrong characterized perceptions so that in the following cycle these perceptions will get the high likelihood for order.
- Also, It allocates the weight to the prepared classifier in every emphasis as per the accuracy of the classifier. The more exact classifier will get high weight.
- This procedure emphasize until the total training information fits with no mistake or until came to the predetermined most extreme number of estimators.
- To arrange, play out a "vote" over the entirety of the learning calculations you assembled.

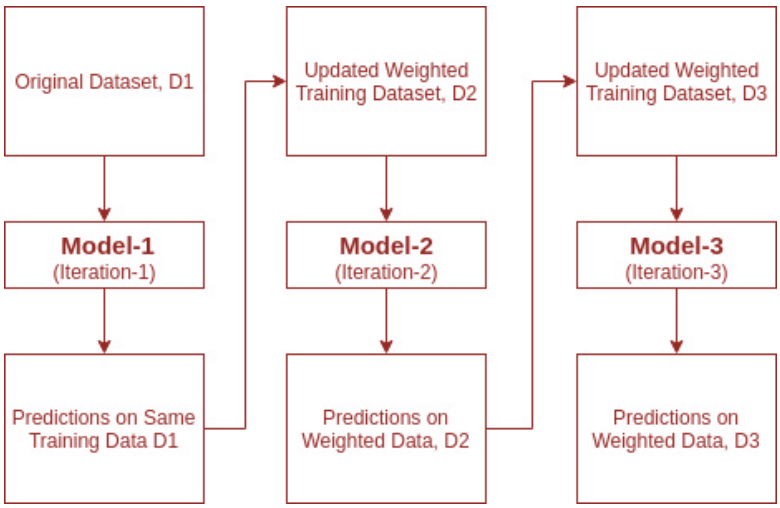


Fig4. AdaBoost

CHAPTER 3 HANDLING IMBALANCED DATASET

3.1. INTRODUCTION

Have you at any point confronted an issue where you have such a little example for the positive class in your dataset that the model can't learn?

In such cases, you get a quite high accuracy just by foreseeing the majority class, however you neglect to catch the minority class, which is frequently the purpose of making the model in any case. Such datasets are a quite regular event and are called as an imbalanced dataset.

Imbalanced datasets are a unique case for classification problem where the class conveyance isn't uniform among the classes. Ordinarily, they are created by two classes: The majority (negative) class and the minority (positive) class. Imbalanced datasets can be found for various use cases in different spaces:

- Financial areas: Fraud identification datasets generally have a fraud rate of ~1–2%
- Clinical: Does a patient has a disease?
- Advertisement Serving: Click expectation datasets additionally don't have a high click-through rate.
- Content control: Does a post contain NSFW content?
- Transportation/Airline: Will Airplane failure happen?

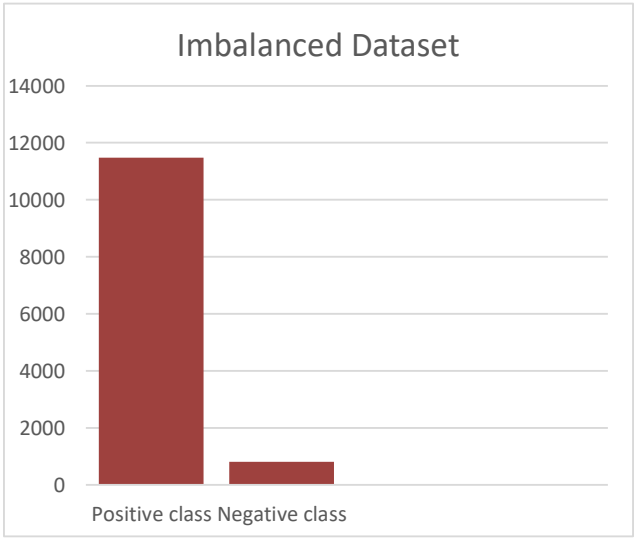


Fig5. Imbalanced Dataset

So, how it effects our model and results. With scarcely any positives comparative with negatives, the training model will invest the greater part of its time in negative models and not gain enough from positive ones. The principle target of adjusting classes is to either expanding the recurrence of the minority class or diminishing the recurrence of the majority class. This is done so as to get roughly a similar number of cases for both the classes.

In this project we are handling the imbalanced dataset using dataset of Google play reviews to classify the sentiment for the working and learning process and results of sampling technique on this dataset.

3.2. SAMPLING TECHNIQUES USED

3.2.1. Under- Sampling Technique

It is a technique where we reduce the sample in the majority class to match up the total length of the minority class sample.

Let's take an example, with 10000 rows with two classes YES and NO with 900 rows of YES class and 100 of the NO class. There is a clear imbalance in the distribution of both. Whatever models will be build would be biased towards the YES sample as the model will be feed with YES sample more than the NO one. To create a balance, random samples from 900 rows of YES class is taken and try to reduce the number of rows from the majority class to get equal to minority class. Best would be to make YES class sample equal to 100 samples, so that both the classes could be weighed equally and then feed to the model. This is how under sampling work. Since there is random exclusion of samples based on many permutation and combination so, there is lots of information loss, discarding useful information which could have been important for building rule classifier. The advantage of using this sampling technique would be improvement in run time and storage problem as the number of training data sample also get reduced if it is huge. We use Random Under-sampling in thus project to deduce the result of under-sampling on our dataset.

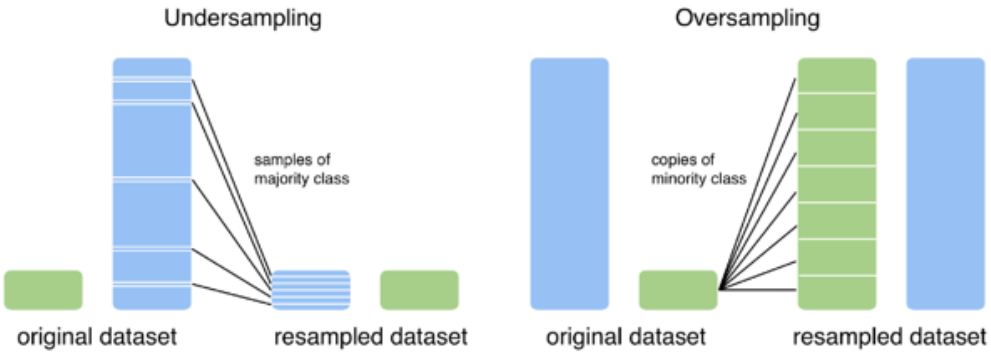


Fig 6. Under Sampling and Over Sampling

322. Over- Sampling Technique

In under sampling, we lose a lot of information be reducing the samples. To tackle that we use over sampling. In over sampling, we kind of increase the sample in the minority class to match up to the number of samples in the majority class. If we have 100 samples in minority class and 900 samples in majority class, then we should increase the minority class sample to reach to length of majority class. This is the basic idea of how over sampling achieved. There are various ways to do it; random over sampling and SMOTE. We are using SMOTE-Synthetic Minority Over Sampling Technique, in this project:

Let's explain it with an example. Suppose there are two features Positive and Negative with Negative as minority class. SMOTE uses nearest neighbor for every point of data by joining them with a line in the minority class dataset and chooses the points which are generated artificially on the lines and takes specified number of synthesized points making minority class samples equivalent to majority class.

So, SMOTE takes a subset of data from the minority class as an example and then new synthetic similar instances are created. These generated instances are then added to the original dataset. Then the new dataset is used as a training sample for the classification models.

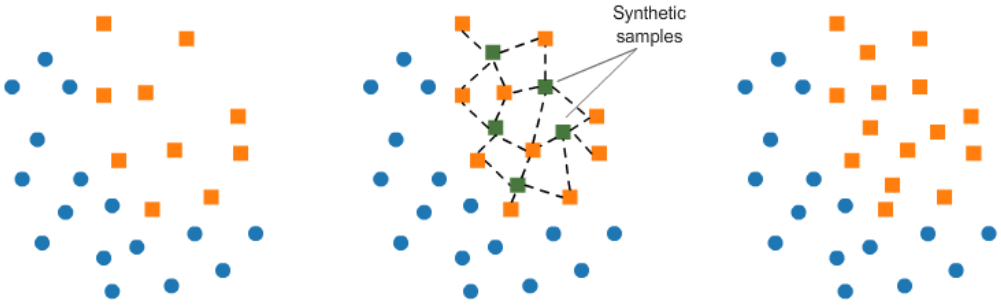


Fig 7. SMOTE Technique

The advantage of using Over Sampling is to that there is no information loss and outperforms Under-sampling. As a limitation, over sampling could lead in the increase of likelihood of overfitting as it replicates the minority class events. Although it does not increase information but raises over- fitting issue causing model to be too specific.

323. Combination of Under-Sampling and Over Sampling

There are blends of oversampling and undersampling strategies that have demonstrated compelling and together might be considered resampling procedures. Two models are the combination of SMOTE with ENN undersampling & SMOTE with Tomek Links undersampling.

SMOTE ENN

SMOTE might be the most well-known oversampling strategy and can be joined with various under sampling strategies. One of the famous under sampling strategy is the Edited Nearest Neighbors, or ENN, rule. This standard includes utilizing k=3 closest neighbors to find those models in a dataset that are misclassified and that are then expelled. It very well may be applied to all classes or simply those models in the majority class.

The mix of SMOTE and ENN is more forceful at down sampling the greater part class than Tomek Links, giving more top to bottom cleaning. They apply the technique, expelling models from both the larger part and minority classes.

ENN is utilized to expel models from the two classes. In this manner, any model that is misclassified by its three closest neighbors is expelled from the preparation set.

SMOTE TOMEK

Tomek Links alludes to a strategy for recognizing sets of closest neighbors in a dataset

that have various classes. Evacuating either of the models in these sets, (for example, the models in the majority class) has the impact of settling on the choice limit in the training dataset less noisy, uproarious or ambiguous.

In particular, first the SMOTE strategy is applied to oversample the minority class to a decent dissemination, at that point models in Tomek Links from the majority classes are distinguished and expelled.

The combination was appeared to give a decrease in False Negative at the expense of an expansion in False Positive for a binary classification.

CHAPTER 4 IMPLEMENTATION AND RESULTS

For this project the dataset that have been taken into account is different Google App Reviews from Playstore. These reviews were reposited to Kaggle.com. So, the dataset .csv file has been downloaded from same. Below Fig 8. shows various google apps whose reviews have been used.

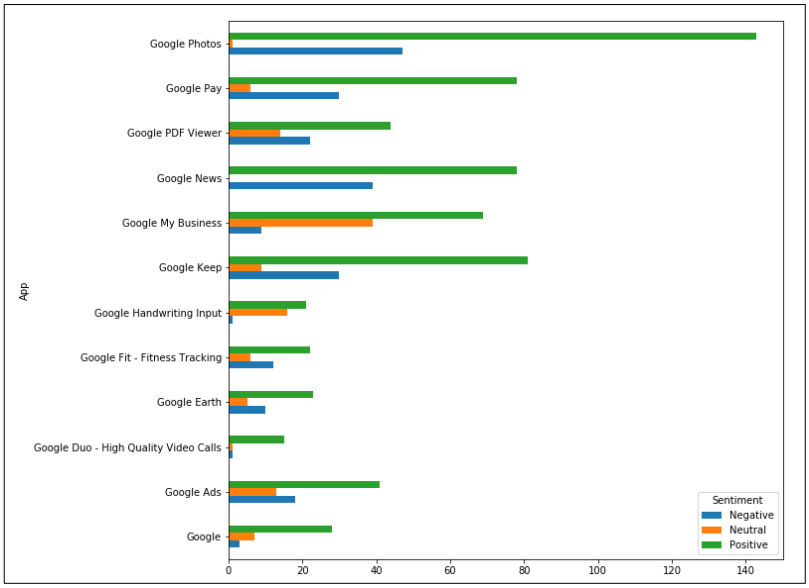


Fig 8. Different Google Apps from Playstore

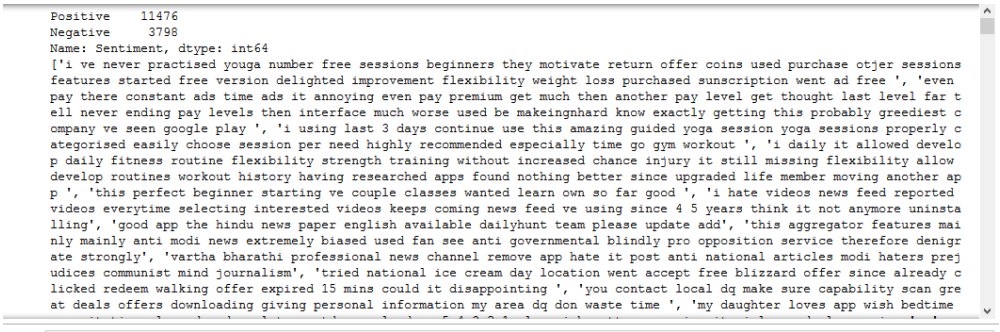
- Figure 9 describes the dataset file in form of table. It contains column of different app names with reviews either Positive or Negative featured as Sentiment column. Further training set use this column to make the ML model grasp the pattern and give best performance.

	App	Translated_Review	Sentiment
0	Daily Yoga - Yoga Fitness Plans	I've never practised youga number free session...	Positive
1	Daily Yoga - Yoga Fitness Plans	Even pay there's constant ads time ads! It's a...	Negative
2	Daily Yoga - Yoga Fitness Plans	I using last 3 days continue use. This amazing...	Positive
3	Daily Yoga - Yoga Fitness Plans	I daily. It allowed develop daily fitness rout...	Positive
4	Daily Yoga - Yoga Fitness Plans	This perfect beginner starting. I've couple cl...	Positive

Fig 9. Detailed view of dataset

The online reviews thus retrieved are then cleaned to get only required stuffs which would efficiently use for feature extraction, as all data is not useful for this work. So, it is done by

removing punctuation marks, special characters, all single characters, multiple spaces, converting from uppercase to lowercase, etc. So Fig 10. depicts the cleaned data ready for feature extraction.



```
Positive    11476
Negative     3798
Name: Sentiment, dtype: int64
['i ve never practised youga number free sessions beginners they motivate return offer coins used purchase otjer sessions
features started free version delighted improvement flexibility weight loss purchased sunscription went ad free ', 'even
pay there constant ads time ads it annoying even pay premium get much then another pay level get thought last level far t
ell never ending pay levels then interface much worse used be makeingnhard know exactly getting this probably greediest c
ompany ve seen google play ', 'i using last 3 days continue use this amazing guided yoga session yoga sessions properly c
ategorised easily choose session per need highly recommended especially time go gym workout ', 'i daily it allowed develo
p daily fitness routine flexibility strength training without increased chance injury it still missing flexibility allow
develop routines workout history having researched apps found nothing better since upgraded life member moving another ap
p ', 'this perfect beginner starting ve couple classes wanted learn own so far good ', 'i hate videos news feed reported
videos everytime selecting interested videos keeps coming news feed ve using since 4 5 years think it not anymore uninsta
lling', 'good app the hindu news paper english available dailyhunt team please update add', 'this aggregator features mai
nly mainly anti modi news extremely biased used fan see anti governmental blindly pro opposition service therefore denigr
ate strongly', 'vartha bharathi professional news channel remove app hate it post anti national articles modi haters prej
udices communist mind journalism', 'tried national ice cream day location went accept free blizzard offer since already c
licked redeem walking offer expired 15 mins could it disappointing ', 'you contact local dq make sure capability scan gre
at deals offers downloading giving personal information my area dq don waste time ', 'my daughter loves app wish bedtime
```

Fig10. Few reviews after cleaning process.

After removing all the unwanted part from our dataset, feature extraction is performed using POS tagging. Below two figures, Fig 11 and Fig 12 shows the wordnet of various words before and after POS tagging. Adjective words are taken into account so as to analyse the sentiments mainly related to this part of speech.

Fig 11 Wordnet before POS tagging

Fig 12 Wordnet after POS tagging

The imbalanced dataset distribution over binary sentiment i.e., Positive and Negative is showed in Fig 13.. With a total of 15274 reviews; 11476 reviews are Positive and 3798 are Negative reviews, making this distribution imbalanced. So the model build on this dataset would be more biased to Positive class and would result new set of data as

positive.

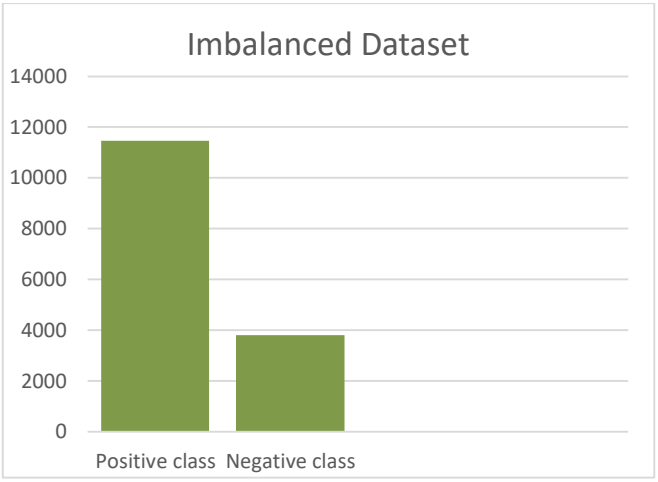


Fig 13. distribution of dataset showing its imbalanced nature

Splitting dataset: Dataset have been split into training set and testing set. These are in 80–20 ratio. Out of 15274 reviews, 12219 were taken as training set and 3055 as testing set.

Applying improper evaluation measurements for model created utilizing imbalanced information can be hazardous. Envision our training data is the one outlined in Fig 13. In the event that accuracy is utilized to gauge the integrity of a model, a model which classifies all testing data into "Positive" will have an excellent accuracy, yet clearly, this model won't give any important data to us. So we used Precision and recall for showing the evaluation.

4.1. PERFORMANCE MEASURES

We have used three evaluation metrics: Precision, Recall, Accuracy.

When a model is built to predict a certain class or category, then a way is needed to measure how accurate the predictions are. Precision, recall and F1 does the same. They measure the classification model’s accuracy. In terms of Confusion Matrix, where we get to know how many times a model correctly or incorrectly predicts a class. Precision, Recall and F1 use these to measure a model as making many mistakes when predicting class, or if it’s doing a pretty good job at being spot on in its predictions. But these three metrics measure different things. Let’s say a classification model predicts or classifies

circle and triangle shape. If the model eludes a lot of mistakes in predicting circles as triangles, then the model is said to have a high precision. In the same way if the model eludes a lot of faults in predicting triangles as circles, then the model has a high recall. A model should be mend to aim high values in both precision and recall, in which the models eludes a lot of mistakes, by predicting both the shapes correctly. But what if the case is, that the model aces the ability to predict one class and fails at predicting the other. So this is when F1 comes in picture. F1 takes into account both precision and recall. F1 scores on balancing both recall and precision. So, if the models predict both the shapes correctly then the model is said to have high F1 score. There are some cases where one would like to focus on either of the precision or recall. Suppose we have class A as "have extreme disease" and class B as "no disease". What if the model predicts class A data as "no disease"? That would be bad prediction. It's okay if the models predict class B data as "have extreme disease". It's better to have a false alarm. So, one would want the model to elude mistaking "have extreme disease" for "no disease", or mistaking A for B. This means one wants to focus on recall. So, coming back to our study, most of the authors have used Precision, recall and F1, where 27 are witnessed or used precision as their performance metrics, 24 studies had recall and F1 as their predicting measure. Most of the studies used all three of them along with accuracy in their paper and their result [14] [15] [16] [17] [18] [19] [20] [21].

Talking about the Confusion Matrix, it's a layout that represents how many predicted classes or categories were correctly predicted & how many were not. It is used to estimate the output of a predicting model with a class outcome to see the number of classes that were predicted correctly as their true class. For this we need to understand TP, TN, FP, FN. See them as, class A correctly anticipated as class A, class B correctly anticipated as class B, class A incorrectly predicted as class B and class B predicted incorrectly as class A. In terms of true and false, target class A is correctly ancipitated as class A which is true, or incorrectly predicted as B in fact was A, which is false. The target class A is the positive and the other class B is our negative, so then a TP and TN is, a positive class A correctly ancipitated as class A and negative class B correctly predicted as B respectively. So the aim would be to get as many numbers of predictions of A and B as possible, aiming for more trues rather than falses. Fig. 14 represents the confusion matrix. It will tell that how many times an actual class a was predicted as B and vice versa. Or if they were correctly classed as their true labels. Studies which followed this metric are [22] [23] [24] [25] [26] [27].

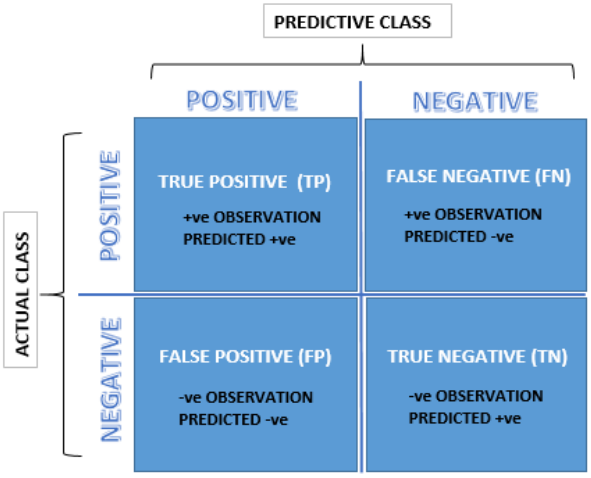


Fig. 14 Confusion Matrix metric

4.2. EXPERIMENTAL RESULTS PERFORMANCE COMPARISONS

Classification techniques are applied on imbalanced dataset. Below table shows the result of all four classifiers for each class; positive and negative. Further this table results will be compared to the tables that would depict the result of sampling techniques applied on these classification techniques after balancing or sampling the dataset by all four methods described earlier in section 3.

Table 1: Results of classification techniques on imbalanced dataset

	CLASS	PRECISION	RECALL	F1-SCORE	ACCURACY SCORE
NAÏVE BAYES	Negative	0.36	0.88	0.51	0.56
	Positive	0.92	0.46	0.61	
DECISION TREE	Negative	0.79	0.82	0.80	0.89
	Positive	0.94	0.93	0.93	
RANDOM FOREST	Negative	0.00	0.00	0.00	0.742
	Positive	0.74	1.00	0.85	
ADABOOST	Negative	0.85	0.67	0.75	0.884
	Positive	0.89	0.96	0.93	

UNDERSAMPLING

Table 2: Results of classification techniques using Under-Sampling

RANDOM UNDER- SAMPLING TECHNIQUE				
	CLASS	PRECISION	RECALL	ACCURACY
NAÏVE BAYES	Negative	0.69	0.83	0.72
	Positive	0.77	0.61	
DECISION TREE	Negative	0.84	0.81	0.82
	Positive	0.81	0.84	
RANDOM FOREST	Negative	0.88	0.90	0.88
	Positive	0.89	0.87	
ADABOOST	Negative	0.83	0.92	0.86
	Positive	0.91	0.81	

OVERSAMPLING:

Table 3: Results of classification techniques using SMOTE

SMOTE				
	CLASS	PRECISION	RECALL	ACCURACY
NAÏVE BAYES	Negative	0.68	0.91	0.74
	Positive	0.87	0.58	
DECISION TREE	Negative	0.92	0.93	0.92
	Positive	0.93	0.93	
RANDOM FOREST	Negative	0.74	0.93	0.80
	Positive	0.91	0.69	
ADABOOST	Negative	0.87	0.90	0.88
	Positive	0.90	0.87	

COMBINATION OF UNDERSAMPLING AND OVERSAMPLING:

Table 4: Results of classification techniques using SMOTE-ENN

SMOTE-ENN				
	CLASS	PRECISION	RECALL	ACCURACY
NAÏVE BAYES	Negative	0.99	0.91	0.92
	Positive	0.73	0.97	
DECISION TREE	Negative	0.99	0.98	0.98
	Positive	0.94	0.98	
RANDOM FOREST	Negative	1.00	0.99	0.99
	Positive	0.98	0.98	
ADABOOST	Negative	0.97	0.99	0.96
	Positive	0.97	0.88	

Table 5: Results of classification techniques using SMOTE-TOMEK

SMOTE-TOMEK				
	CLASS	PRECISION	RECALL	ACCURACY
NAÏVE BAYES	Negative	0.67	0.91	0.73
	Positive	0.87	0.58	
DECISION TREE	Negative	0.92	0.93	0.92
	Positive	0.93	0.92	
RANDOM FOREST	Negative	0.95	0.98	0.96
	Positive	0.98	0.95	
ADABOOST	Negative	0.86	0.92	0.89
	Positive	0.92	0.86	

Above all four table describes the result of all the sampling techniques used to balance the dataset and then build the model based on those balanced set. Fig.15 shows the combination result of accuracy score of each machine learning model for every individual sampling technique. Though there is increase in the value of each precision , recall and F1- score value of sampling technique results as compared to the table showing result without sampling technique (Table 1).

Table: 6. Comparison result of Sampling techniques on ML techniques

	CLASS	Random Under-Sampling	SMOTE	SMOTE ENN	SMOTE Tomek
NAÏVE BAYES	Negative	0.75	0.78	0.95	0.77
	Positive	0.68	0.70	0.83	0.69
DECISION TREE	Negative	0.82	0.93	0.99	0.92
	Positive	0.82	0.93	0.96	0.93
RANDOM FOREST	Negative	0.89	0.83	0.99	0.96
	Positive	0.88	0.79	0.98	0.96
ADABOOST	Negative	0.87	0.89	0.98	0.89
	Positive	0.85	0.89	0.92	0.89

Table 6 represents the comparison between all the sampling techniques that are applied to all four ML technique, NB, RF, DT, AdaBoost. As compared to Tabl 1. All the values of all three metrics show increase in value. And among four sample techniques, SMOTE ENN shows best result than other three sampling technique.

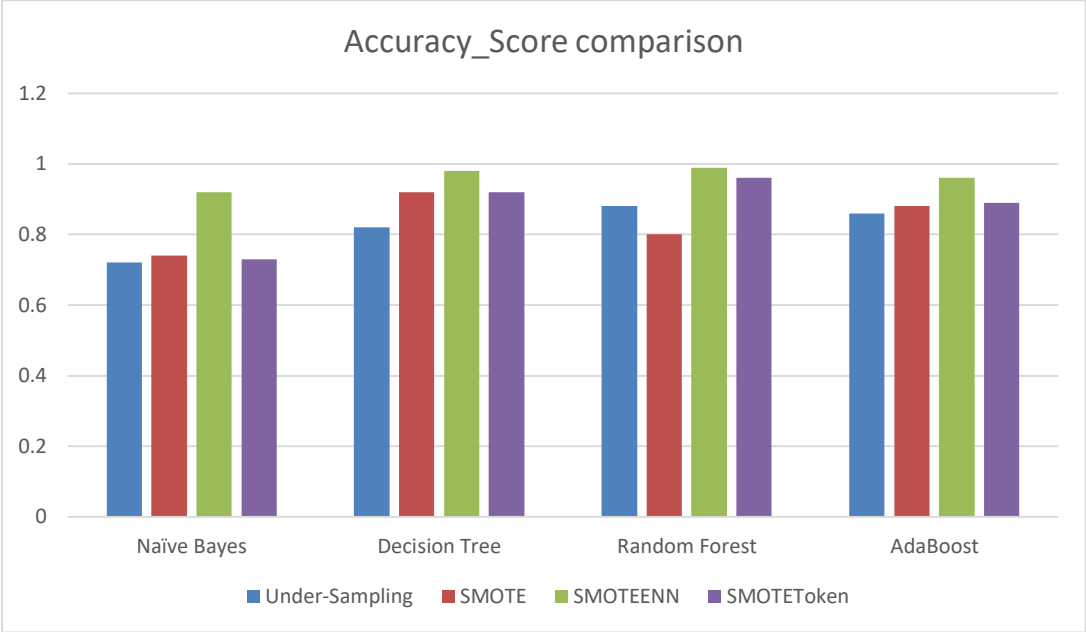


Fig. 15. Accuracy Score for all classifier

The main focus would be the variation in precision Recall and F1-score of all the results after applying each sampling technique. It has been improved in all the sampling technique result as compared to the Table 1. And among all the sampling technique SMOTE ENN can be seen winning.

CHAPTER 5 CONCLUSION AND FUTURE WORK

In this chapter, we first briefly summarize the main work in the thesis. And then gather the findings and make some comments on them. At last, we suggest possible future work in order to better tackle the problem.

5.1 SUMMARIZATION & CONCLUSION

Our aim in this thesis project is to handle the imbalanced dataset and train the classification models to perform better and give efficient result. So, summarizing the whole work. Chapter 1 introduces the objective of pursuing this project and and what results are we aiming for. Along with that the detail of the organization of the report section wise.

In Chapter 2, the background knowledge that is rewuired before handling the imbalanced dataset is focused in this section. With SA introduction and its need in todays time and what are the fields that could use this analysis. Also the working model of SA is discussed along with its technique. Next subsection 2.2. deals with related work handling the same objective that many authors have done using ML techniques and various sampling techniques. These papers have been surveyed and summarized in this section. Along with that in section 2.3., various ML techniques that is going to be used in this work is discussed; focusing on NB,DT,RF,AdaBoost.

In Chapter 3, The main aim of the paper, dealing with the imbalanced dataset is discussed; its introduction, why imbalanced datset will be a problem and how it can be rectified by adding value to the ML techniques for classification. The main four techniques Under-sampling, Oversampling(SMOTE) and combination of both done by SMOTEENN and SMOTE Tomek, their working has been discussed.

In Chapter 4, the the whole implementation part is discussed. All the process and results have been carried out using Python Scikit Learn library and Imblearn library on Jupyter Notebook. Dataset have been taken from Kaggle.com related to google play reviews of only the google apps to analyse the sentiment. The performance metric used in this work is Accuracy, Precision, Recall and F1 Score. Table 1, table 2, table 3, Table 4, Table 5 shows the result of all the work and comparison between the sampling techniques used

on imbalanced dataset giving results for all four ML techniques. So, with all the technique of sampling used; SMOTE ENN showed the best results. Random Under-sampling did not give and efficient result as compared to results on Table 1. (results on imbalanced dataset without sampling technique).

5.2 FUTURE SCOPE

Considering other big data, comparing different data sampling algorithms and investigating other effective solutions to class imbalance in big data. Other would be to advance the robustness and efficiency of producing sampling rate. There are many methods of sampling that are done by combining more than one technique. So, different hybrid approach should be considered. Also to cope with the imbalanced dataset, various ensemble techniques are used that works on classification models by automatically making the balanced result of the model. Also, different real world application scenario can be focused on where there is huge need of balancing the data to get an appropriate result.

APPENDICES

APPENDIX 1: LIST OF PUBLICATIONS (PUBLISHED)

2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)

Sentiment Analysis on Product Reviews

Rahul
Department of Computer Science and Engineering
Delhi Technological University
Delhi, 110042, India
rahul@dtu.ac.in

Vasundhara Raj
Department of Computer Science and Engineering
Delhi Technological University
Delhi, 110042, India
vasundhara.raj94@gmail.com

Monika
Department of Computer Science and Engineering
Delhi Technological University
Delhi, 110042, India
monika.siwaliya@gmail.com

***Abstract*— Sentiment Analysis is a process of analyzing and categorizing the emotion or sentiment over any given review or text piece in order to know what the reviewer wants to express in the form of positive, negative or neutral. Today, people are highly interested in buying things online from any e-commerce site or they search for a product review in order to know the quality and one's perception toward that product before buying. The product provider also gets to know about the user's opinion over a product. This can help the company to improve its marketing strategy and quality of product in their favor. Sentiment analysis uses various semantic approaches like on these online reviews to extract as much feature it can and categorize the type of opinion .Some techniques also help in rating the product value based on user's opinion. This paper is a literature survey including various authors and their sentiment techniques on product or online review.**

Keywords— Sentiment Analysis, Online Reviews, Opinion Extraction, Polarity, SentiWordNet, etc.

I. INTRODUCTION

Online buying today is the most preferred and trending way for buying any product like clothes, home essentials, electronic gadgets, sport equipments, books, etc. For buying a particular thing, a user needs to be sure about the quality of the product, as only product pictures and description can be viewed online. So there is a review system for the product, reviewed by the users who bought it. They write about the product quality they received and other opinions over that product bought by the. This review helps other user buying the similar product. And these review sets can be taken from different platforms such as Amazon, Flipkart, Ebay, taobao, yelp, etc.

Suppose a review says, "This product is awesome. Good to use." It can be seen that the words like awesome and good has been used. These words indicate a positive opinion. But to mine the sentiments technically these reviews undergo assorted process and classification techniques. Firstly, these reviews are gathered and proper noise removal is done in order to process the texts. Then different extraction techniques like Total Weighted Score Computing Method, Neutral/Polar/Irrelevant Classification Model, Weighing and Aggregation Scheme, Intrinsic and Extrinsic Domain Relevance Approach, etc can be used to extract the feature of the pre-processed text. After that classification techniques are used to classify the sentiments into positive, negative or neutral followed by polarity detection. These sentiments are determined over various words, phrases. Final step is to verify the precision of the result obtained.

Different techniques used in classification of sentiments on different types of datasets from different platforms. These techniques help to detect polarity of the text or review used. Sentiment Analysis techniques includes machine learning and lexical based approach. This paper includes collection of many research papers briefing their techniques used in sentiment analysis classification, their observation on their techniques used. Techniques like Support Vector Machine, Naïve Bayes, Neural Network, etc. are highly used in this area. There are many more techniques that work good on their datasets.

The rest of the paper is structured as follows. Section II introduces about sentiment analysis process model and techniques used in it. Section III describes the various techniques used by different authors to accomplish the classification of sentiment and perform these techniques on different review sets. In section IV finally, we conclude the paper.

II. SENTIMENT ANALYSIS PROCESS MODEL AND TECHNIQUES

This Section of paper consists of Sentiment Analysis Process Model and different approaches used in it to analyze the sentiment.

A. Process Model of Sentiment Analysis

Sentiment Analysis is a process to determine the person's opinion or attitude over any product, service or organization. We will discuss each step in brief as depicted in Fig 1.

ISBN: 978-1-7281-4826-7/19/$31.00 ©2019 IEEE

APPENDIX 2: ACCEPTED and PUBLISHED in Journal of Critical Reviews (SCOPUS)

Online Reviews Over Sentiment Analysis using Machine Learning: A Systematic Review

Rahul, Vasundhara Raj, Delhi Technological University, Delhi, 110042, India
rahul@dtu.ac.in, vasundhara.raj94@gmail.com

Abstract— One of the key territories of NLP is Sentiment Analysis, the capacity to comprehend emotional tones in speech and text. This Systematic Literature Review has focused on papers between 2015 to 2020, taken from trusted and credible database such as IEEE Xplore, Science Direct and Springer. A total of 70 papers have been chosen for this review. This SLR approach is followed to get an effective insight on various work being done in this research field using Machine learning techniques: supervised or unsupervised. Different research questions have been looked up and discussed. The result shows that most of the work have used SVM for classification techniques and accuracy as the performance metrics. Also most of the dataset are yielded from e-commerce sites for product reviews, reviews in form of tweets from twitter and in various other fields like hospitality reviews, movie reviews and other social networking sites opinions.

INTRODUCTION

TODAY with the approach of Machine Learning procedures, Artificial Intelligence and the expanded mechanization in each division from the IOT based smart gadgets to modern robots, the need is being acknowledged in each other part which has not recovered the true abilities of this innovation. The development of these ideas can be acknowledged from the way that robots were prior just characterized by a mechanical arm and now it has moved to instruct pendant strategies to now even programmable robots. One of the key zones of natural language processing (NLP), a subset of AI is Sentiment Analysis (SA), the capacity to comprehend passionate tones in discourse and print. It is a territory that is the concentration for various diverse utilitarian applications.

A. Why there is a need of SA?

Let's understand it by taking an application of customer service. Success of a company and its service growth directly depends on its client. If the client enjoys the product, then its success of the company else it needs to improvise it by making some changes in it. So in order to know whether customer like your product or not you need to analyse them. One of the attributes is to analyse the customers by their sentiment. That's how the sentiment comes into the picture. So, a procedure of computationally recognizing and categorizing sentiments from a text, and conclude whether the outlook of the writer for the specific topic or the product is positive, negative or neutral. As a user, one does not perform a SA but one does look for response like before purchasing a product one looks into what other customer have to say about that product whether it is decent or bad. And that user analyses it manually. Now, while considering it at a company level, that how will a company analyse what a customer thinks about their service. A company have customer in abundance and so their reviews. This is where a company needs to carry out SA to know if the product or service is actually working in the market or not. Similarly goes for other applications like Companies running hotels can analyse its service by left out online reviews by customer detailing how was their stay or service at the hotel. Twitter comments section contains a lot of opinions and those can be used to analyse over an issue or topic categorizing polarity and predicting the overall sentiment. Sentiment Analysis can also be used in marketing businesses in call centres too. A customer service bot can be improvised to handle various sentiment of the customer. As a customer can be frustrated or happy by the service or using any slang words or those words which can be treated as both negative and positive in different context, for example: "Not Bad!!".

REFERENCES

[1] Rahul, V. Raj, and Monika, "Sentiment Analysis on Product Reviews," 2019 Int. Conf. Comput. Commun. Intell. Syst. Sentim., pp. 5–9, 2020, doi: 10.1109/icccis48478.2019.8974527.

[2] T. Lu, Y. Huang, W. Zhao, and J. Zhang, "The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN," Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2019, pp. 370–374, 2019, doi: 10.1109/ICCSNT47585.2019.8962430.

[3] I. Benchaji, S. Douzi, and B. El Ouahidi, "Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection," Lect. Notes Networks Syst., vol. 66, pp. 220–229, 2019, doi: 10.1007/978-3-030-11914-0_24.

[4] Y. Pang, Z. Chen, L. Peng, K. Ma, C. Zhao, and K. Ji, "A signature-based assistant random oversampling method for malware detection," Proc. - 2019 18th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. IEEE Int. Conf. Big Data Sci. Eng. Trust. 2019, pp. 256–263, 2019, doi: 10.1109/TrustCom/BigDataSE.2019.00042.

[5] F. Yin, Q. Du, M. Chen, Q. Bao, and Y. Gao, "Fault Prediction for Network Equipment based on Oversampling Method in Imbalanced Dataset," 2019 IEEE Int. Conf. Consum. Electron. - Taiwan, ICCE-TW 2019, pp. 2019–2020, 2019, doi: 10.1109/ICCE-TW46550.2019.8991912.

[6] S. Zhang, X. Shang, W. Wang, and X. Huang, "Optimizing the classification accuracy of imbalanced dataset based on SVM," ICCASM 2010 - 2010 Int. Conf. Comput. Appl. Syst. Model. Proc., vol. 4, no. Iccasm, pp. 338–341, 2010, doi: 10.1109/ICCASM.2010.5620370.

[7] P. Shukla and K. Bhowmick, "To improve classification of imbalanced datasets," Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICIIECS.2017.8276044.

[8] S. Harjai, S. K. Khatri, and G. Singh, "Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique," 2019 4th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2019, pp. 123–128, 2019, doi: 10.1109/ISCON47742.2019.9036162.

[9] A. Puri and M. K. Gupta, "Comparative Analysis of Resampling Techniques under Noisy Imbalanced Datasets," IEEE Int. Conf. Issues Challenges Intell. Comput. Tech. ICICT 2019, 2019, doi: 10.1109/ICICT46931.2019.8977650.

[10] T. Hasanin, T. M. Khoshgoftaar, J. Leevy, and N. Seliya, “Investigating random undersampling and feature selection on bioinformatics big data,” Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol., pp. 346–356, 2019, doi: 10.1109/BigDataService.2019.00063.

[11] H. Zhang, Z. Li, H. Shahriar, L. Tao, P. Bhattacharya, and Y. Qian, “Improving prediction accuracy for logistic regression on imbalanced datasets,” Proc. - Int. Comput. Softw. Appl. Conf., vol. 1, pp. 918–919, 2019, doi: 10.1109/COMPSAC.2019.00140.

[12] A. Hanskunatai, “A New Hybrid Sampling Approach for Classification of Imbalanced Datasets,” 2018 3rd Int. Conf. Comput. Commun. Syst. ICCCS 2018, pp. 278–281, 2018, doi: 10.1109/CCOMS.2018.8463228.

[13] S. Choirunnisa and J. Lianto, “for Handling Imbalanced Data,” 2018 Int. Semin. Res. Inf. Technol. Intell. Syst., pp. 276–280, 2017.

[14] A. Mukwazvure and K. P. Supreethi, “A hybrid approach to sentiment analysis of news comments,” 2015 4th Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2015, 2015, doi: 10.1109/ICRITO.2015.7359282.

[15] A. Salinca, “Business Reviews Classification Using Sentiment Analysis,” Proc. - 17th Int. Symp. Symb. Numer. Algorithms Sci. Comput. SYNASC 2015, pp. 247–250, 2016, doi: 10.1109/SYNASC.2015.46.

[16] T. Dholpuria, Y. K. Rana, and C. Agrawal, “A sentiment analysis approach through deep learning for a movie review,” Proc. - 2018 8th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2018, pp. 173–181, 2018, doi: 10.1109/CSNT.2018.8820260.

[17] J. Jabbar, I. Urooj, W. Junsheng, and N. Azeem, “Real-time sentiment analysis on E-Commerce application,” Proc. 2019 IEEE 16th Int. Conf. Networking, Sens. Control. ICNSC 2019, pp. 391–396, 2019, doi: 10.1109/ICNSC.2019.8743331.

[18] Q. Umer, H. Liu, and Y. Sultan, “Sentiment based approval prediction for enhancement reports,” J. Syst. Softw., vol. 155, pp. 57–69, 2019, doi: 10.1016/j.jss.2019.05.026.

[19] A. Gelbukh, “Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015 Cairo, Egypt, April 14-20, 2015 Proceedings, Part II,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9042, pp. 114–125, 2015, doi: 10.1007/978-3-319-18117-2.

[20] J. Singvejakul, C. Chaiboonsri, and S. Sriboonchitta, The Dependence Structure and Portfolio Optimization in Economic Cycles : An Metadata of the chapter that will be visualized in SpringerLink, vol. 2, no. March. Springer International Publishing, 2019.

[21] Z. Rahimi, S. Noferesti, and M. Shamsfard, “Applying data mining and machine learning techniques for sentiment shifter identification,” Lang. Resour. Eval., vol. 53, no. 2, pp. 279–302, 2019, doi: 10.1007/s10579-018-9432-0.

[22] R. Joshi and R. Tekchandani, “Comparative analysis of twitter data using supervised classifiers,” Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016, vol. 2016, 2016, doi: 10.1109/INVENTIVE.2016.7830089.

[23] A. M. Alayba, V. Palade, M. England, and R. Iqbal, “Arabic language sentiment analysis on health services,” 2017 IEEE Int. Work. Arab. Scr. Anal. Recognit. Arab., pp. 114–118, 2017, doi: 10.1109/asar.2017.8067771.

[24] Y. Gupta and P. Kumar, “Real-Time Sentiment Analysis of Tweets: A Case Study of Punjab Elections,” Proc. 2019 3rd IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2019, pp. 1–12, 2019, doi: 10.1109/ICECCT.2019.8869203.

[25] A. Alrehili and K. Albalawi, “Sentiment analysis of customer reviews using ensemble method,” 2019 Int. Conf. Comput. Inf. Sci. ICCIS 2019, pp. 1–6, 2019, doi: 10.1109/ICCISci.2019.8716454.

[26] G. Vinodhini and R. M. Chandrasekaran, “A comparative performance evaluation of neural network based approach for sentiment classification of online reviews,” J. King Saud Univ. - Comput. Inf. Sci., vol. 28, no. 1, pp. 2–12, 2016, doi: 10.1016/j.jksuci.2014.03.024.

[27] W. C. T. and J. C. M. S. Arrieta Rodriguez Eugenia, Francisco Edna Estrada, “Advances in Artificial Intelligence - IBERAMIA 2016,” vol. 10022, no. November, pp. 259–70, 2016, doi: 10.1007/978-3-319-47955-2.