A Project Report

on

FAKE REVIEW DETECTION
Using PassiveAggressiveClassifier

Submitted in partial fulfillment of the
requirement for the award of the degree of

Bachelor Of Technology in Computer Science and Engineering

Under The Supervision of
K. ANANDHAN
Assistant Professor
Department of Computer Science and Engineering

Submitted By

MANISH KUMAR 19SCSE1010208
AMRESH KUMAR 19SCSE1010630

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
DECEMBER, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **" FAKE REVIEW DETECTION USING PassiveAggressiveClassifier "** in partial fulfillment of the requirements for the award of the **Bachelor Of Technology** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of **Mr. K. Anandhan**, Department of Computer Science and Engineering, Galgotias University, Greater Noida.

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

Manish Kumar 19SCSE1010208

Amresh Kumar 19SCSE1010630

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Mr. K. Anandhan
AP/SCSE
Galgotias University

CERTIFICATE

The Project Viva-Voce examination of **Manish Kumar: 19SCSE1010208, Amresh Kumar: 19SCSE1010630** has been held on _____ and their work is recommended for the award of Bachelor Of Technology.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021

Place: Greater Noida

ACKNOWLEDGMENT

I am sincerely thankful to Galgotias University, Greater Noida for providing me with the opportunity to write a research paper on the topic **“ FAKE REVIEW DETECTION USING PassiveAggressiveClassifier ”.**

We would like to thank our guide **Mr. K. Anandhan** for guiding us in every stage of this project. Without his support it would have been very difficult for me to prepare such a meaningful and interesting project.

This paper has helped me a lot to learn about machine learning and PassiveAggressiveClassifiers. I hope it helps people to have a basic understanding of fake reviews and the need for tackling the same.

ABSTRACT

In the recent year we have been experiencing a huge surge on internet as many people have started using internet as we have seen decline in the price of internet in most of the countries as a result we can also see nowadays that there have been many paid and fake reviews flooding the e-commerce websites like Amazon, Flipkart and many other e-commerce websites based on which many Customers make decisions based on these fake reviews or comments provided by others who have had similar experiences. In today's competitive environment, anyone may write anything, which has resulted in an increase in the amount of bogus reviews. So to overcome this problem of fake reviews we will be using a machine learning model. The model will be trained with many datasets available on Kaggle and other sites. Machine learning is a subtopic of artificial intelligence which basically uses the provided data in this field and predicts the output based on the experience. Furthermore we can test the model using testing data which is also a part of the data set. We are using Passive Active Classifier which results in an accuracy of 89.32%.

Keywords : Machine Learning, Passive Active Classifier, Model

ACRONYMS

ML	Machine Learning
PAC	Passive Aggressive Classifier
UL	Unsupervised learning
CM	Confusion Matrix
SL	Supervised learning
AI	Artificial intelligence

TABLE OF CONTENTS

List of Table

List of Figures

CHAPTER - 1

INTRODUCTION

1.1 Introduction

Nowadays, everyone has access to smartphones and other devices on which one can access the internet. Due to this boom use of social media and e-commerce sites, it is found that around 90% of people in the 30-50yrs are considering the reviews posted on social media and e-commerce sites for buying some product or consideration for some business or travelling. Not only that, nowadays even younger people are more likely to consider reviews for buying some product. And these reviews can have a huge impact on customer as well as business

For example, there has been many researches in which it was found that many of the reviews which are being posted on the e-commerce sites are either fake or not posted by genuine customers. There have been many cases out there in the news in which many big businesses are buying these fake reviews by paying the fake reviewers some incentives for posting these fake positive reviews which will obviously increase the sales and gain them fame. And on the other hand one can buy these fake reviewers who are also called as opinion spammers and the process of posting fake review is called as the opinion spamming to either accelerate their financial gain or for defaming the products of other sellers or business services so it is the reason why these fake positive or negative reviews can be devastating for both customers as well as business.

In past these cases of fake reviews have become common which are even high profile case reported on daily news channels and there have also been many cases regarding Amazon and Flipkart who are pointed for invoking this act and not taking proper measures regarding this and many a times they have been summoned by the Competition Commission of India (CCI) for not taking proper action.

But nowadays these big giants have been taking many steps and using many machine learning algorithms to filter these fake reviews and they are using methods for recommending users

whether to consider some review or ignore those reviews and also reviewers are given tags whether they are genuine buyers or not.

The main algorithms used in the recent past have been using **Supervised Machine Learning** but during the first few years of using this algorithm there have many obstacles due to inadequate availability of the review data but as the time passed there have been many techniques like using human labelled data for better accuracy of predicting whether a review is fake or genuine. But there were also some drawbacks of using the human labelled data because these data just contained so much noise in them.

Machine learning is the field of study that allows computers to learn without being explicitly programmed. Using machine learning, we don't need to provide explicit instructions to Computers for reacting to some special situations. We need to provide training to the computers to find real-time solutions for the specific problems. The chess game is a famous example where machine learning is being used to play chess. The code lets the machine learn and optimizes itself over repeated games.

Machine Learning is broadly classified into two main categories.

- Supervised Machine Learning
- Unsupervised Machine Learning
- Supervised Learning

Supervised learning is similar to having a trainer or teacher who supervises all the machines' reactions and tells step-by-step solutions to specific problems. It's like a hand-holding way of teaching the computers what to do. One real-time example of supervised learning is recognizing different types of images using computers. Being humans, we also learn by this model as we are taught to recognize different objects like a car by repeat exposures. In the same way, machines are taught.

We feed a different set of some specific images into a machine where each image has a specific identifier to identify the type of the image. However, computers are taught so that every time the

particular blend of pixels comes in front of the computer; it can recognize the type of image loaded into the model dataset.

Supervised learning works in a way that the computer can learn through the previous exposures; for example, if a computer sees a car object and recognizes it like a car, then next time, it should be able to identify any different image of car object by identifying a lot of features that are similar to previously identified images of Car.

When we train a machine learning model for image recognition, we present many images where every single image is attached to a label so that the data can be clearly labeled and gets stored in a machine learning model. Once we complete the training, we should present an object's image that is not part of the training data, and the machine should be able to identify it by classifying all its previous learnings.

This is the most fundamental type of Supervised learning, which is called Classification. Out machine learning model must be able to classify the different bunch of images. It must be programmed so that the different objects can be recognized according to their unique characteristics. However, we can create a generic classifier so that it is not dependent on learning data. We don't need to recode the entire model on changing the training data.

Unsupervised Learning

In Supervised learning, the specific kind of dataset is loaded into computers to learn through the repeat exposures to the dataset. In this section of the article, we will discuss Unsupervised learning. It is one of the other major types of machine learning. Instead of providing training data where every piece of data is clearly labeled, we provide the unstructured training data in unsupervised learning. We want the model to sense the dataset so that it learns to find the structure in unstructured data.

In other words, we can say, in unsupervised learning, we don't tell computers the kind of data. Instead, we want the computers to see the structure in the data by observing which way the data is being organized.

One type of Unsupervised learning is called clustering, in which the computer looks at the dataset and its features and can figure out the separate clusters in which the data is maintained.

Reinforcement Learning

We have covered supervised learning in which we have loaded the labeled training data in the machine learning models so that the computers can classify the data and perform regressions to identify the dataset. We have also covered Unsupervised learning. We have loaded the unstructured unlabeled dataset grouped in separate clusters, and we want computers to be smart enough to identify the separate clusters.

In this section of the article, we will discuss Reinforcement Learning. As humans, we are much experienced in reinforcement learning. We tend to learn through reinforcement. For example, if driving through a route that is full of traffic jam, we'll ignore going through the same route on other days. There are two kinds of reinforcements we generally come through, 1. Positive, 2. Negative Reinforcement.

The same way machines work in the case of Reinforcement Algorithms. One of the real-time examples of reinforcement learning is a Chess game where the Computer with the Reinforcement learning algorithm calculates the winning probability with every move. The computer might come through positive as well negative reinforcement with every single move. However, through many cycles of training and by practicing more and more games, the computers will learn about which moves in which situations will lead to an increase in its winning percentage.

Passive Aggressive Classifiers

The Passive-Aggressive algorithms are a family of Machine learning algorithms that are not very well known by beginners and even intermediate Machine Learning enthusiasts. However, they can be very useful and efficient for certain applications.

Note: This is a high-level overview of the algorithm explaining how it works and when to use it. It does not go deep into the mathematics of how it works.

Passive-Aggressive algorithms are generally used for large-scale learning. It is one of the few **'online-learning algorithms'**. In online machine learning algorithms, the input data comes in sequential order and the machine learning model is updated step-by-step, as opposed to batch learning, where the entire training dataset is used at once. This is very useful in situations where there is a huge amount of data and it is computationally infeasible to train the entire dataset because of the sheer size of the data. We can simply say that an online-learning algorithm will get a training example, update the classifier, and then throw away the example.

A very good example of this would be to detect fake news on a social media website like Twitter, where new data is being added every second. To dynamically read data from Twitter continuously, the data would be huge, and using an online-learning algorithm would be ideal.

Passive-Aggressive algorithms are somewhat similar to a Perceptron model, in the sense that they do not require a learning rate. However, they do include a regularization parameter.

How Passive-Aggressive Algorithms Work:

Passive-Aggressive algorithms are called so because :

- **Passive:** If the prediction is correct, keep the model and do not make any changes. i.e., the data in the example is not enough to cause any changes in the model.
- **Aggressive:** If the prediction is incorrect, make changes to the model. i.e., some change to the model may correct it.

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time. There's a good chance you've interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chatbots, and other consumer conveniences. But NLP also plays a growing role in enterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-critical business processes.

Confusion matrix

Confusion matrix is a very popular measure used while solving classification problems. It can be applied to binary classification as well as for multiclass classification problems.

Confusion matrices represent counts from predicted and actual values. The output "TN" stands for True Negative which shows the number of negative examples classified accurately. Similarly, "TP" stands for True Positive which indicates the number of positive examples classified accurately. The term "FP" shows False Positive value, i.e., the number of actual negative examples classified as positive; and "FN" means a False Negative value which is the number of actual positive examples classified as negative. One of the most commonly used metrics while performing classification is accuracy. The accuracy of a model (through a confusion matrix) is calculated using the given formula below.

Accuracy can be misleading if used with imbalanced datasets, and therefore there are other metrics based on confusion matrix which can be useful for evaluating performance. In Python, confusion matrix can be obtained using "confusion_matrix()" function which is a part of "sklearn" library [17]. This function can be imported into Python using "from sklearn.metrics

import confusion_matrix." To obtain confusion matrix, users need to provide actual values and predicted values to the function

The confusion matrix consists of four basic characteristics (numbers) that are used to define the measurement metrics of the classifier. These four numbers are:

1.**TP** (True Positive): TP represents the number of patients who have been properly classified to have malignant nodes, meaning they have the disease.
2.**TN** (True Negative): TN represents the number of correctly classified patients who are healthy.
3.**FP** (False Positive): FP represents the number of misclassified patients with the disease but actually they are healthy. FP is also known as a Type I error.
4.**FN** (False Negative): FN represents the number of patients misclassified as healthy but actually they are suffering from the disease. FN is also known as a Type II error.

Performance metrics of an algorithm are accuracy, precision, recall, and F1 score, which are calculated on the basis of the above-stated TP, TN, FP, and FN.

Accuracy of an algorithm is represented as the ratio of correctly classified patients (TP+TN) to the total number of patients (TP+TN+FP+FN).

Precision of an algorithm is represented as the ratio of correctly classified patients with the disease (TP) to the total patients predicted to have the disease (TP+FP).

Recall metric is defined as the ratio of correctly classified diseased patients (TP) divided by total number of patients who actually have the disease.
The perception behind recall is how many patients have been classified as having the disease. Recall is also called sensitivity.

F1 score is also known as the F Measure. The F1 score states the equilibrium between the precision and the recall.

1.2 OBJECTIVE

We are trying to create an environment where the user doesn't need to get influenced by the fake reviews or rumors rather than detect them by themselves. The main purpose of this project is to create a tool for detecting language patterns and categorize them into real or fake reviews with the help of machine learning and natural language . We have trained algorithms that work on the particular type of domain and do not provide any results when exposed to articles from other domains. Since articles from different sources have different structures and formats. So it is very difficult to train an algorithm that works best on all review domains.

1.3 RESEARCH PROBLEM

It could be very much important to the companies and even customers who are solely relying on these reviews posted and it could financially affect the companies because as we can see many people take decision based on opinions and reviews as we have also seen from many research that there are many opinion spammers being hired to boost sales and also to defame other brand or seller's products so as to gain more reach to the customers who are iterating on these e-commerce sites because many of e-commerce websites the algorithms to promote the products with better reviews.

The data can be extracted from Kaggle which is well known for data availability and also most of the websites like Flipkart, Amazon, and other ecommerce sites have a huge repository of review data which could also be used for training and testing data sets which would be more effective.

Our method will be using content and the word used in the reviews as the original customer's mindsets are different when writing reviews then that of fake customers writing reviews, word choice pattern can be used for segregation of fake and truthful reviews.

1.4 Formulation of Problem

As the use of online products and applications has been increasing at a rapid rate, the competition within the market is increasing day by day. Business officials so as to frame their products and defame other competitor products may get people from other marketplace or hire them to post and provide fake judgements on the products. So to safeguard the authenticity of the net products and opinions various steps and methods are needed to be applied.

It could be very much important to the companies and even customers who are solely relying on these reviews posted and it could financially affect the companies because as we can see many people take decision based on opinions and reviews as we have also seen from many research that there are many opinion spammers being hired to boost sales and also to defame other brand or seller's products so as to gain more reach to the customers who are iterating on these e-commerce sites because many of e-commerce websites the algorithms to promote the products with better reviews.

So these acts of opinion spammers could hamper the decision of the customers and sales of the companies who are selling genuine products but are bring notices due to usage of the fake reviews as there have been many cases that big brands becomes obstacle for the upcoming brands with good products as they don't want to lose their market share.

To tackle this problem the best and the most effective method could be using supervised machine learning as this field has already proven its capability in detecting fake reviews and many other fake things with usage of proper algorithms as well as using proper training and testing data set. The data can be extracted from Kaggle which is well known for data availability and also most of the websites like Flipkart, Amazon, and other ecommerce sites have a huge repository of review data which could also be used for training and testing data sets which would be more effective. Our method will be using content and the word used in the reviews as the original customer's mindsets are different when writing reviews then that of fake customers writing reviews, word choice pattern can be used for segregation of fake and truthful reviews.

1.4.1 Tools and technologies used

Tools Used:

Hardware Requirements:

- Processor: Pentium or later processors (2.4 GHz).
- Hard Disk: 10 GB.
- Ram: 4GB or more

Software Requirements:

- Operating system: Windows 7 or later versions.
- Coding Language: Python
- IDE: Python ide (Google Colaboratory).

Technologies Used:

- Supervised ML
  - ➔ Passive Aggressive Classifier
- Natural Language Processing

SOFTWARE USED

Google Colaboratory : Google colab is a data analysis tool which combines code output, data and descriptive documents in a single collaborative document.

Colab is a free notebook environment that runs entirely in the cloud. It lets you and your team members edit documents, the way you work with Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.

Google is quite aggressive in AI research. Over many years, Google developed an AI framework called TensorFlow and a development tool called Colaboratory. Today TensorFlow is open-sourced and since 2017, Google made Colaboratory free for public use. Colaboratory is now known as Google Colab or simply Colab.

LANGUAGE USED

- Python
- Machine learning (concept).

LIBRARIES USED

- **numpy** : For Computing array

  NumPy stands for Numerical Python. It is a Python library used for working with an array. In Python, we use the list for the purpose of the array but it's slow to process. NumPy array is a powerful N-dimensional array object and its use in linear algebra, Fourier transform, and random number capabilities. It provides an array object much faster than traditional Python lists.

  Types of Array:

  A. One Dimensional Array
  B. Multi-Dimensional Array

- **pandas** : For reading the csv and and convert into dataframe

  Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance & productivity for users.

- **sklearn** : For shuffling the data, performance metrics.

  Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

CHAPTER - 2

LITERATURE SURVEY

- In a recent study a method was proposed by E.I Elmurngi and A. Gherbi, Journal of Computer Science, 2018, DOI: 10.3844/jcssp.2018.714.726, using an open source software tool called 'Weka tool' to implement machine learning algorithms using sentiment analysis to classify fair and unfair reviews from amazon reviews based on three different categories: positive, negative and neutral words. In this research work, the spam reviews are identified by only including the helpfulness votes voted by the customers along with the rating, deviations are considered which limits the overall performance of the system. Also, as per the researcher's observations and experimental results, the existing system uses Naive Bayes classifier for spam and nonspam classification where the accuracy is quite low which may not provide accurate results for the user.

- Initially N. O'Brien and J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, 2019, DOI: 10.1056/NEJMoa1911303 ,have proposed solutions that depend only on the features used in the data set with the use of different machine learning algorithms in detecting fake news on social media. Though different machine learning algorithms the approach lacks in showing how accurate the results are. B. Wagh,J.V.Shinde, P.A.Kale worked on twitter to analyze the tweets posted by users using sentiment analysis to classify twitter tweets into positive and negative. They made use of K-Nearest Neighbour as a strategy to allot them sentiment labels by 7 training and testing the set using feature vectors. But the applicability of their approach to other types of data has not been validated.

- B. Wagh,J.V.Shinde, P.A.Kale worked on twitter to analyze the tweets posted by users using sentiment analysis to classify twitter tweets into positive and negative.They made use of K-Nearest Neighbor as a strategy to allot them sentiment labels by training and testing the set using feature vectors. But the applicability of their approach to other type of data has not been validated

RELATED WORK

In the last few years, depending on the context, researchers have proposed many different approaches to tackle the issue of the assessment of the credibility of the information diffused through social media. Historically, the concept of credibility has been in turn associated with believability, trustworthiness, perceived reliability, expertise, accuracy, and with numerous other concepts or combinations of them. According to Fogg and Tseng , credibility is a perceived quality of the information receiver, and it is composed of multiple dimensions. Different characteristics can be connected to: (i) the source of information, (ii) the information itself, i.e., its structure and its content, and (iii) the media used to disseminate information .

It has been demonstrated that, when considering these characteristics in terms of credibility, the impact of the delivery medium can change the perception that people have about sources of information and information itself. For this reason, one important question to be tackled nowadays is whether new media in the digital realm introduce new factors that may concur to credibility assessment. In the Social Web, evaluating information credibility deals with the analysis of the user-generated content, the authors' characteristics, and the intrinsic nature of social media platforms, i.e., the social relationships connecting the involved entities.

These characteristics, namely features, can be simple linguistic features associated with the text of the UGC, they can be additional meta-data features associated for example with the content of a review or a tweet, they can also be extracted from the behavior of the users in social media, i.e., behavioral features, or they can be connected to the user profile (if available).

Furthermore, different approaches have taken into consideration product-based features, in the case of review sites where products and/or services are reviewed, or have considered social features, which exploit the network structure and the relationships connecting entities in social media platforms. In the last years, several approaches have been proposed to assess in an automatic or semi-automatic way the credibility of information in the Social Web; in particular, the most investigated tasks have been the identification of: (i) opinion spam in review sites , (ii) fake news in microblogging sites , and (iii) potentially harmful/inaccurate online health

information. In general, the majority of these approaches focus on data-driven techniques, which classify UGC with respect to credibility by employing different models. With regard to opinion spam detection, and in particular to fake review detection, which is the focus of this paper, the approaches that have produced the best results are generally based on supervised or semi-supervised machine learning techniques that take into account both review- and reviewer centric features. The first approaches were purely linguistic, in the sense that they employed simple textual features extracted from the text of reviews, often in the form of unigrams and/or bigrams. Other linguistic approaches have proposed generative classifiers based on language models. It has been demonstrated by Mukherjee et al. in that focusing only on linguistic features is not effective to detect fake reviews from real datasets, since it is practically impossible for a human reader to distinguish between credible and not credible information by simply reading it, especially in an era where the skills in writing false reviews are constantly improving.

For this reason, more effective multi-feature-based approaches have been proposed, which employ several features of different nature in addition to simple linguistic ones, either by applying supervised or semi-supervised machine learning , or by implementing the Multi-Criteria Decision Making (MCDM) paradigm. These approaches usually focus on small labeled datasets for evaluation purposes, constituted in most of the cases by ‘near ground truth’ data.

They usually avoid considering features that are extracted from the social ties constituting the network of entities (e.g., users, products, reviews) considered by the review site. On the contrary, this kind of feature is often utilized (together with the other features previously described) by graph-based approaches. These latter approaches are in most of the cases unsupervised, even if sometimes they can be coupled with a supervised learning phase on a limited number of classification labels. With respect to supervised approaches, totally unsupervised solutions generally provide slightly worst results.

This paper, by considering the effectiveness of supervised solutions, discusses and analyzes on a general level the most appropriate review- and reviewer-centric features that have been proposed so far in the literature to detect fake reviews; moreover, it proposes some new features suitable for this aim, in particular to detect singleton fake reviews, an issue that has not yet received the

attention it deserves. To avoid the problem of the limited size of the labeled datasets considered up to now by the literature, two large-scale publicly-available datasets presented in have been employed for evaluation purposes.

FEATURE ANALYSIS

As briefly introduced in Section II, many and different are the features that have been considered so far in the review site context to identify fake reviews. In some cases, features belonging to different classes have been considered separately by distinct approaches. In other cases, the employed features constitute a subset of the entire set of features that could be taken into account; furthermore, new additional features can be proposed and analyzed to tackle open issues not yet considered, for example the detection of singleton fake reviews. For these reasons, in this section we provide a global overview of the various features that can be employed to detect fake reviews. Both significant features taken from the literature and new features proposed in this article are considered. Since the most effective approaches discussed in the literature are in general supervised and consider review- and reviewer-centric features, these two classes will be presented in the following sections. The choices behind the selection of the features belonging to the above mentioned classes will be detailed along each section. When the features are taken from the literature, they will be directly referred to the original paper where they have been initially proposed. The absence of the reference will denote those features that have been widely used by almost every proposed technique.

Finally, the presence of the label denoted by [new] will indicate a feature proposed for the first time in this article. A. Review Centric Features The first class of features that have been considered, is constituted by those related to a review. They can be extracted both from the text constituting the review, i.e., textual features, and from meta-data connected to a review, i.e., metadata features. In every review site, the time information regarding the publication of the review, and the rating (within some numerical interval) about the reviewed business are metadata, are always provided. In addition, in relation to metadata features, those connected to the cardinality of the reviews written by a given user must be carefully studied. In fact, a large part of reviews are singletons, i.e., there is only one review written by a given reviewer in a certain period of time (this means that in the user account there is only one review at the time of the analysis). For this kind of reviews, specific features must be designed. In fact, as it will be illustrated in the following, many of the features that have been proposed in the literature are

based on some statistics over several reviews written by the same reviewer. In the case of singletons, these features loose their relevance in assessing credibility.

Therefore, the definition of suitable features that are effective for detecting also singleton fake reviews becomes crucial.

1)Textual Features: as briefly illustrated in Section II, it is practically impossible to distinguish between fake and genuine reviews by only reading their content. The analysis provided by Mukherjee et al. in has shown that the KL-divergence between the languages employed by spammers and non spammers in Yelp is very subtle. However, the good results obtained by using linguistic features on a domain specific dataset (i.e., a Yelp's dataset containing only New York japanese restaurants), show that at least on a domain specific level, textual features can be useful. It is possible to use Natural Language Processing techniques to extract simple features from the text, and to use as features some statistics and some sentiment estimations connected to the use of the words.

• Text: several approaches employ as textual features both unigrams and bigrams extracted from the text of reviews,as illustrated.

• Text statistics: several statistics on the review content have been proposed as features by Li et al. in :

– Number of words, i.e., the length of the review in terms of words;

– Ratio of capital letters, i.e., the number of words containing capital letters with respect to the total number of words in the review;

– Ratio of capital words, i.e., considering the words where all the letters are uppercase;

– Ratio of first person pronouns, e.g., 'I', 'mine', 'my',etc.;

– Ratio of 'exclamation' sentences, i.e., ending with the symbol '!'.

• Sentiment evaluations:

– Subjectivity, i.e., a number representing the proportion of subjective words (expressing sentiment, judgment) as opposed to

objective (descriptive) words.

2) Meta-data Features: these kinds of features are extracted from the meta-data connected to reviews, or they can be generated by reasoning on the reviews' cardinality with respect to the reviewer and the entity reviewed.

• Basic features:

– Rating, i.e., the rating attributed in the review to the entity, in the form of some numerical value belonging to a given interval

(e.g., 1-5 'stars');

– Rating deviation, i.e., the deviation of the evaluation provided in the review with respect to the entity's average rating;

– Singleton, i.e., it indicates the fact that the review is the only one provided by a reviewer in a given period of time (e.g., a day).

These basic features rely on some simple and intuitive heuristics. A fake review tends to contain a more 'extreme' rating with respect to genuine reviews, thus implying that the rating deviation from the entity's average rating is higher; furthermore, a singleton review provided by a user could indicate that s/he is not particularly involved in the review site community, which constitutes a possible indication of unreliability.

• Burst features: it is said that reviews for an entity are 'bursty' when there is a sudden concentration of reviews in a time period. These review bursts can be either due to sudden popularity of the entities reviewed or to spam attacks. Since it has been proven that reviews in the same burst tend to have the same nature, it is possible to easily identify groups of fake

reviews by analyzing the nature of the burst. Two burst detection studies have been described in. Taking inspiration from the just cited works, in this paper several features considering burstiness have been introduced. These features are related to the time window in which a review has been posted, relatively to a given reviewed entity. Basically, a review is more likely to be fake if it is posted on a day when the number of reviews is abnormally high, and when the average rating associated with an entity in a review (in a specific time window) varies significantly with respect to the entity's average rating (in general it decreases, for example passing from 3.5/5 to 2/5).

CHAPTER - 3

WORKING OF PROJECT

3.1 PROPOSED SOLUTION

To tackle this problem the best and the most effective method could be using supervised machine learning as this field has already proven its capability in detecting fake reviews and many other fake things with usage of proper algorithms as well as using proper training and testing data set.

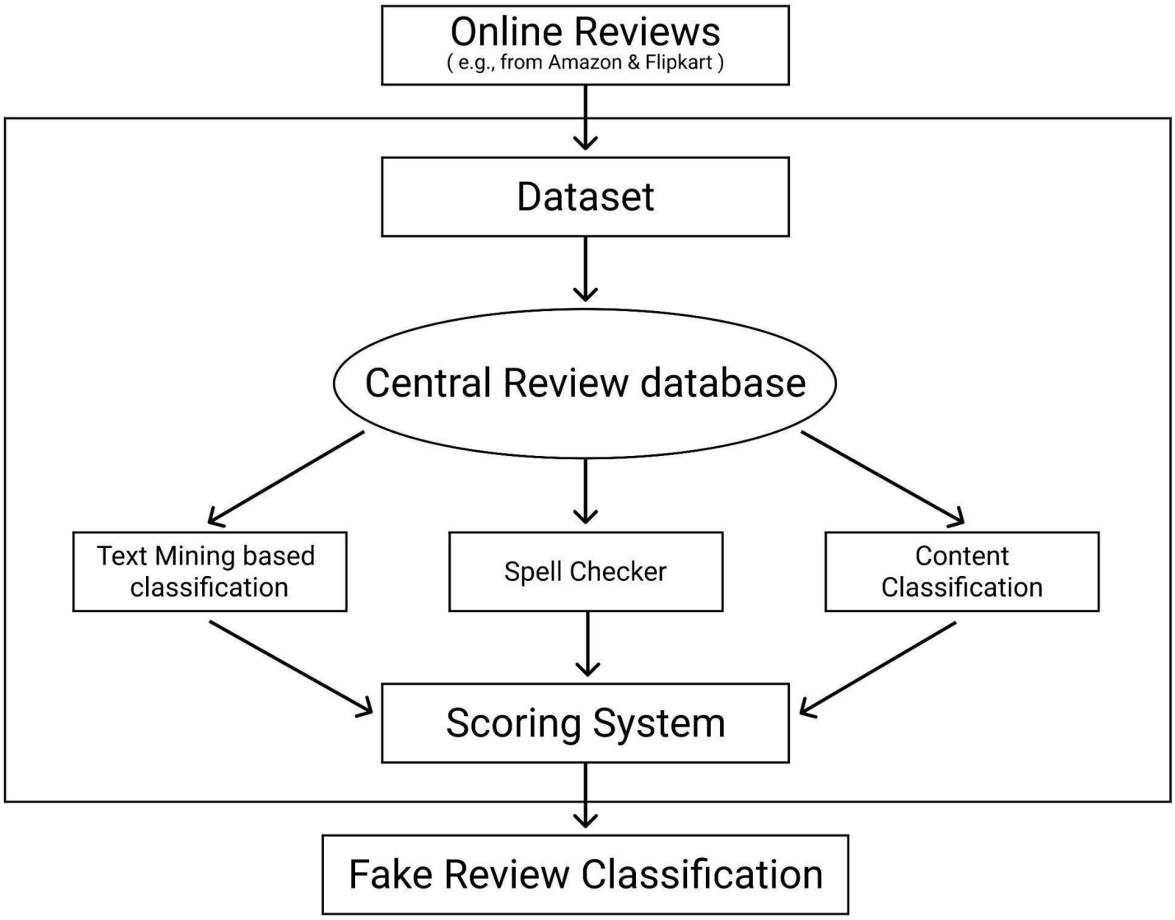


Fig. I. System Diagram

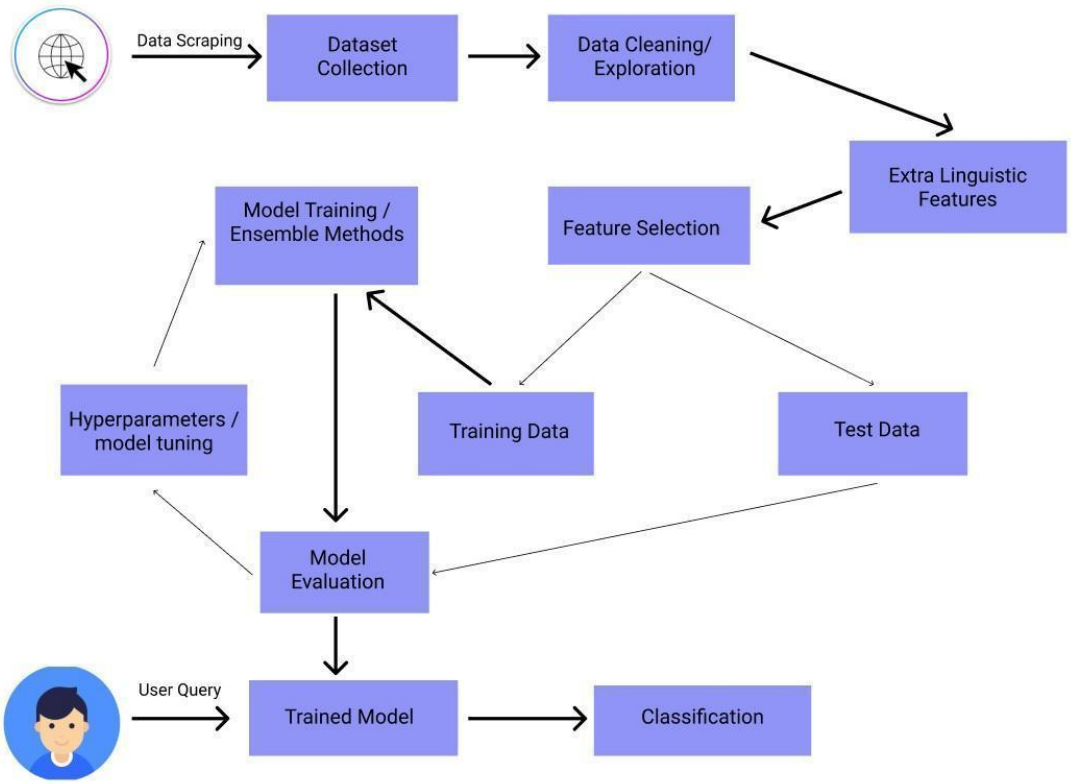


Fig. II. Block Diagram of the Proposed System

Data-oriented: It points to different forms of fake reviews data, such as data collection, language and psychological consent of fake reviews .

Feature-oriented: Its main aim is to explore the functional features for detecting fake reviews from many data sources, such as review content and social factors.

Model-oriented: It opens the gateway to build more practical and potent models for fake reviews detection, that includes supervised, semi-supervised and unsupervised models.

Application-oriented: It bounds research that goes after fake reviews detection, such as fake reviews diffusion and intervention.

Step 1: Feature Extraction.

There are different structures of review, useful review is extracted from them and written in the form of review content which is listed below.

- **Source**:- Author or Publisher of the review article.
- **A headline**:-Short title that gives little idea about the review.
- **Body Text**:-Main a text of the review that elaborates the details of the review story.
- **Image/Video**:-Visual representation of the review.

Step 2: Model Construction.

Dataset:-review can be collected from different sources such as review websites, review agency homepages, search engines, and social media websites. However, determining the correctness of review manually is a difficult task, usually requiring good analysts with an expert in their domain who performs very careful analysis of claims and additional evidence of the review to categorize them in real or fake, context, and reports from authoritative sources.

Evaluation Metrics: In this we have evaluated the performance of algorithms that are written to detect fake reviews detection problems, various metrics are used . In this section, we review the most used metrics for the detection of fake reviews . Most existing approaches consider the fake reviews problem as a classification problem that predicts whether a review is fake or real:

- True Positive (TP): It is used when fake reviews are predicted and they are actually described as fake reviews.
- True Negative (TN): It is used when true review is predicted and they are actually described as true review.
- False Negative (FN): It is used when true reviews are predicted and they are actually described as fake reviews.

- False Positive (FP): It is used when fake reviews are predicted and they are actually described as true reviews.

1. Precision=|T P||T P|+|F P|

2. Recall=|T P||T P|+|F N|

3. F1 = 2·Precision Recall Precision +Recall

4. Accuracy=|T P|+|T N||T P|+|TN|+|F P|+|F N|

CHAPTER - 4

Result and Conclusion

RESULT

▾ FAKE REVIEW DETECTION

```
[1] #Mount Google drive
    from google.colab import drive
    drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
[2] #Change the directory to the path where you have all your input files
    %cd /content/drive/My Drive/UNIVERSITY/PROJECT_III
    #Below command shows you all thee files present in your current directory
    # !ls
```

```
/content/drive/My Drive/UNIVERSITY/PROJECT_III
```

```
[3] #Importing libraries
    import numpy as np
    import pandas as pd
    import itertools
```

1. Mounted Google drive.
2. Changed the path to the location of the files.
3. Imported the libraries that we will be using.

```
[4] #importing dataset
    dataset=pd.read_csv("FakeReview.csv")
```

```
dataset.shape
dataset.head(10)
```

	category	rating	label	text_
0	Home_and_Kitchen_5	5.0	CG	Love this! Well made, sturdy, and very comfor...
1	Home_and_Kitchen_5	5.0	CG	love it, a great upgrade from the original. I...
2	Home_and_Kitchen_5	5.0	CG	This pillow saved my back. I love the look and...
3	Home_and_Kitchen_5	1.0	CG	Missing information on how to use it, but it i...
4	Home_and_Kitchen_5	5.0	CG	Very nice set. Good quality. We have had the s...
5	Home_and_Kitchen_5	3.0	CG	I WANTED DIFFERENT FLAVORS BUT THEY ARE NOT.
6	Home_and_Kitchen_5	5.0	CG	They are the perfect touch for me and the only...
7	Home_and_Kitchen_5	3.0	CG	These done fit well and look great. I love th...
8	Home_and_Kitchen_5	5.0	CG	Great big numbers & easy to read, the only thi...
9	Home_and_Kitchen_5	5.0	CG	My son loves this comforter and it is very wel...

4. Imported the Dataset.
5. Printed the head(first 10 lines) of the Dataset.

```
[6] #checking labels
    labels=dataset.label
    labels.head(10)
```

```
0    CG
1    CG
2    CG
3    CG
4    CG
5    CG
6    CG
7    CG
8    CG
9    CG
Name: label, dtype: object
```

```
[7] dataset.replace(to_replace="CG",value="FAKE",inplace=True)
    dataset.replace(to_replace="OR",value="REAL",inplace=True)
```

6. Checking the labeled Dataset which was categorized into two parts - CG and OR which are then replaced by FAKE and REAL using replace function.
7. splitting the dataset into training and testing dataset using sklearn.model_selection function train_test_spit.

```
[8] #checking labels
    labels=dataset.label
    labels.head(10)
```

```
0    FAKE
1    FAKE
2    FAKE
3    FAKE
4    FAKE
5    FAKE
6    FAKE
7    FAKE
8    FAKE
9    FAKE
Name: label, dtype: object
```

```
[9] #training dataset
    from sklearn.model_selection import train_test_split
    x_train,x_test,y_train,y_test=train_test_split(dataset['text_'], labels, test_size=0.2, random_state=7)
```

```
[10] from sklearn.feature_extraction.text import TfidfVectorizer
     tfidf_vectorizer=TfidfVectorizer(stop_words='english',max_df=0.7)
     tfidf_train=tfidf_vectorizer.fit_transform(x_train)
     tfidf_test=tfidf_vectorizer.transform(x_test)
```

8. Using tf_idf_vectorizer for finding the importance of a particular text or word in that particular sentence.

9. Importing PassiveAggressiveClassifier for predicting the output.
10. Training the classifier using the test data and finding the accuracy using the test data .

```
[11] #DataFlair - Initialize a PassiveAggressiveClassifier
     from sklearn.linear_model import PassiveAggressiveClassifier
     from sklearn.metrics import accuracy_score, confusion_matrix
     pac=PassiveAggressiveClassifier(max_iter=50)
     pac.fit(tfidf_train,y_train)

     #DataFlair - Predict on the test set and calculate accuracy
     y_pred=pac.predict(tfidf_test)
     score=accuracy_score(y_test,y_pred)
     print(f'Accuracy: {round(score*100,2)}%')
```

```
Accuracy: 83.53%
```

```
[12] confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])
```

```
array([[3415,  530],
       [ 802, 3340]])
```

11. Number of iterations for training is defined as 50 which will make the model to repeatedly learn from the data for 50 times.

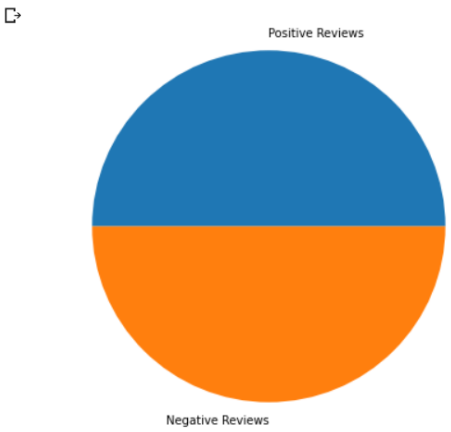
12. plotting the confusion matrix.

```
[13] from matplotlib import pyplot as plt
     pos = 0
     neg = 0
     for score in dataset['label']:
         if score == "REAL":
             pos+=1
         elif score == "FAKE":
             neg += 1

     #Visualiing the distribution of Sentiment
     values = [pos, neg]
     label = ['Positive Reviews', 'Negative Reviews']

     fig = plt.figure(figsize =(10, 7))
     plt.pie(values, labels = label)
     print(pos,"POSITIVE REVIEW")
     print(neg,"NEGATIVE REVIEW")

     plt.show()
```



13. Finding a number of FAKE and REAL reviews and plotting the pie chart for the same.

```
[14] from sklearn.metrics import classification_report
```

```
[15] print(classification_report(y_test,y_pred))
```

```
                precision    recall  f1-score   support

          FAKE       0.81      0.87      0.84      3945
          REAL       0.86      0.81      0.83      4142

      accuracy                           0.84      8087
     macro avg       0.84      0.84      0.84      8087
  weighted avg       0.84      0.84      0.84      8087
```

14. Importing classification_report from sklearn.metrics
15. Printing classification report for the actual and predicted review.

CONCLUSION

It is obvious that reviews play a crucial role in people's decisions. Thus, fake reviews detection is a vivid and ongoing research area. In this paper, a machine learning fake reviews detection approach is presented. In the proposed approach, both the features of the reviews and the behavioral features of the reviewers are considered. The Center For Open Science dataset is used to evaluate the proposed approach. Passive Aggressive classifier is implemented in the developed approach. The results reveal that Passive Aggressive classifier is good in the fake reviews detection process. Also, the results show that considering the behavioral features of the reviewers increases the accuracy. Not all reviewers behavioral features have been taken into consideration in the current work. Future work may consider including other behavioral features such as features that depend on the frequent times the reviewers do the reviews, the time reviewers take to complete reviews, and how frequent they are submitting positive or negative reviews. It is highly expected that considering more behavioral features will enhance the performance of the presented fake reviews detection approach we have filtered the labeled dataset. We have used algorithms like tfidf_vectorizer and PassiveAggressiveClassifier. We have used PassiveAggressiveClassifier to detect the accuracy of our model which is **83.92 percent**.

Also, the approach provides the user with a functionality to recommend the most truthful reviews to enable the purchaser to make decisions about the product. Various factors such as adding new vectors like ratings, emojis, verified purchase have affected the accuracy of classifying the data better.

Future Work

1. To use a real time/ time based datasets which will allow us to compare the user's timestamps of the reviews to find if a certain user is posting too many reviews in a short period of time.

2. To use and compare other machine learning algorithms like logistic regression to extend the research to deep learning techniques.

3. To develop a similar process for unsupervised learning for unlabeled data to detect fake reviews

REFERENCES

[1] R. Barbado, O. Araque, and C. A. Iglesias, “A framework for fake review detection in online consumer electronics retailers,” Information Processing & Management, vol. 56, no. 4, pp. 1234 – 1244, 2019.

[2] S. Tadelis, “The economics of reputation and feedback systems in ecommerce marketplaces,” IEEE Internet Computing, vol. 20, no. 1, pp. 12–19, 2016.

[3] M. J. H. Mughal, “Data mining: Web data mining techniques, tools and algorithms: An overview,” Information Retrieval, vol. 9, no. 6, 2018.

[4] C. C. Aggarwal, “Opinion mining and sentiment analysis,” in Machine Learning for Text. Springer, 2018, pp. 413–434.

[5] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “What yelp fake review filter might be doing?” in Seventh international AAAI conference on weblogs and social media, 2013.

[6] N. Jindal and B. Liu, “Review spam detection,” in Proceedings of the 16th International Conference on World Wide Web, ser. WWW ’07, 2007.

[7] E. Elmurngi and A. Gherbi, Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques. IARIA/DATA ANALYTICS, 2017.

[8] V. Singh, R. Piryani, A. Uddin, and P. Waila, “Sentiment analysis of movie reviews and blog posts,” in Advance Computing Conference (IACC), 2013, pp. 893–898.

[9] A. Molla, Y. Biadgie, and K.-A. Sohn, “Detecting Negative Deceptive Opinion from Tweets.” in International Conference on Mobile and Wireless Technology. Singapore: Springer, 2017.

[10] S. Shojaee et al., “Detecting deceptive reviews using lexical and syntactic features.” 2013.

[11] Y. Ren and D. Ji, “Neural networks for deceptive opinion spam detection: An empirical study,” Information Sciences, vol. 385, pp. 213– 224, 2017.

[12] H. Li et al., “Spotting fake reviews via collective positive-unlabeled learning.” 2014.

[13] N. Jindal and B. Liu, “Opinion spam and analysis,” in Proceedings of the 2008 International Conference on Web Search and Data Mining, ser. WSDM ’08, 2008, pp. 219–230.

[14] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, “What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews,” Journal of Management Information Systems, vol. 33, no. 2, pp. 456–481, 2016.

[15] E. D. Wahyuni and A. Djunaidy, “Fake review detection from a product review using modified method of iterative computation framework.” 2016.