



T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Information Technologies

**CATEGORIES DOCUMENT
USING NAİVE BAYES**

Assalah Abdulwahab Non

Master Thesis

Supervisor

Asst. Prof. Dr. Sefer Kurnaz

Istanbul, 2020

CATEGORIES DOCUMENT USING NAÏVE BAYES

by

Assalah Abdulwahab Non

Information Technologies

Submitted to the Institute of Graduate Studies in partial
fulfillment of the requirements for the degree of Master of
Science

ALTINBAŞ UNIVERSITY

2020

The thesis titled “**CATEGORIES DOCUMENT USING NAIVE BAYES**” prepared and presented by Assalah Abdulwahab NON was accepted as a Master of Science Thesis in Information Technologies.

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Thesis Defense Jury Members:

Asst. Prof. Dr. Sefer KURNAZ

School of Engineering and

Natural Sciences,

Altinbas University

Assoc. Prof. Dr.Yasa EKŞIOĞLU OZOK

School of Engineering and

Natural Sciences,

Altinbas University

Asst. Prof. Dr.Zeynep ALTAN

Faculty of Engineering and

Architecture,

Beykent University

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Approval Date of Institute of Graduate Studies:

___/___/___

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Assalah Abdulwahab Non



DEDICATION

Firstly, thanks for God, I dedicate this thesis to my family, colleagues, and all my friends for their support and encouragement throughout my life and studies. my dedication goes to the assistant supervisor. Professor Dr. Seif Kurtz, For my mother, for her continued encouragement and her support to me also For the soul of my late father For my brothers and my family. Last but not least, a special thanks to the institution that stood by my side and helped me a lot to finish my thesis, the International Academy for Leadership and Development (IALD) And a big thanks to its president, Mr. Abdul Rahman Al-Ahmad And Mr. Mouath Salem Al-Rawi.



ACKNOWLEDGEMENTS

I might want to offer my thanks to every one of the individuals who have upheld me all through the regularly extended periods of this voyage. I might want to thank my advisor, Asst. Prof. Dr. Sefer kurnaz for being my compass notwithstanding when I believed I was lost and being in extraordinary part in charge of the zenith of this work. I might likewise want to thank my supervisors for their supportive exhortation, which incredibly enhanced the nature of this work. And I want to present a big thank to Mr. Dhafar Hamed and his support. Finally, I thank this institution for hosting me during these years, securely earning its place as my home. I am very grateful to my mother and father for my siblings and my family for their unlimited love and for my country, Iraq, which I hope will override all the many tribulations and crisis. All the love.

ABSTRACT

CATEGORIES DOCUMENT USING NAÏVE BAYES

Non, Assalah Abdulwahab,

M.S., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Sefer Kurnaz

Date: June,2020

Pages: 55

Classification plays a crucial role in data mining in which it divides data in rather, separate and separate areas and all area denotes a category. thy are several types of workbooks used to successfully complete the classification task. In addition, classification is restricted only if text documents are classified. The motives of the thesis are to evaluate the classification of multi-layer documents and to know how to achieve accurate classification in the case of text documents. Analysis of moods has an important effect on most types of humanitarian activity and has a great influence on our behaviour. With the evolution of technology and the employ of online technologies, a lot of knowledge has appeared that reflects the opinions of users in many aspects, such as entertainment, politics, business, sports, and technology. This article uses Naive Bayesian (NB) to analysis opinions by studying and classifying text categories in the corresponding chapter. It examines the effectiveness of employ feature extraction, that is, Team Frequency (TF) methods with naive Bayesian (Gaussian and polynomial) classifiers on the accuracy of the classification of BBC documents. accuracy, retrieval, and the F1 number of valid predictions will not estimate the execution of application categories. Conclusion reveal that the use of TF improved accuracy to 97%. It is best suited for our sample accessory.

Keywords: Naïve Bayes, Gaussian, Multinomial, Sentiment Analysis, Term Frequency, Data Set, Pre-processing, Classification Document, Machine Learning.

TABLE OF CONTENTS

	<u>Pages</u>
ABSTRACT	vii
LIST OF TABLES	xii
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS.....	xiv
1. INTRODUCTION.....	1
1.1 GENERAL VIEW	1
1.2 RESEARCH BASICS.....	4
1.2.1 Definition of Main Concept.....	4
1.2.2 Task	4
1.2.3 Techniques.....	4
1.2.4 Natural Language Processing And Information Retrieval In Sentiment Analysis.....	4
1.2.5 Applications of Text Categorization.....	5
1.3 RESEARCH PROBLEM	5
1.4 RESEARCH OBJECTIVES	6
1.5 THE RELATED WORKS	6
1.6 CONTRIBUTION	9
1.7 THESIS ORGNIZATION.....	14
2. THEORETICAL BACKGROUND.....	13
2.1 INTRODUCTION.....	14

2.2 TEXT MINING	13
2.3 BBC (BRITISH BROADCASTING CORPORATION)	14
2.4 SENTIMENT ANALYSIS	13
2.4.1 Approaches for Sentiment Analysis	14
2.4.1.1 Lexicon-based approaches	14
2.4.1.2 Corpus-based.....	15
2.5 PREPROCESSING OF DOCUMENTS	15
2.5.1 Tokenization.....	16
2.5.2 Normalization.....	17
2.5.3 Stop Words.....	17
2.5.4 Stemming	18
2.6 FEATURE EXTRACTION WITH TF	18
2.7 MACHINE LEARNING	19
2.8 DOCUMENTS CATEGORIZATION	21
2.8.1 Training Phase.....	21
2.8.2 Testing Phase.....	21
2.9 DOCUMENTS CATEGORIZATION TYPES.....	22
2.9.1 Hierarchal and Flat Categorization	22
2.9.2 Multi-Label and Single-Label Categorization.....	22
2.9.3 Hard and Soft Categorization	23
2.10 NAÏVE BAYES CLASSIFICATION ALGORITHM.....	23

2.10.1 Gaussian Naive Bayes.....	24
2.10.2 Multinomial Naive Bayes.....	25
2.11 SUMMARY OF THIS CHAPTER.....	26
3. THE PROPOSED SYSTEM	27
3.1 INTRODUCTION.....	27
3.2 THE PROPOSED SYSTEM.....	27
3.2.1 Downloading Data	29
3.2.2 Dataset Description.....	29
3.2.3 Preparing Label.....	29
3.2.4 Preprocessing	29
3.2.4.1 Tokenization.....	30
3.2.4.2 Normalization.....	30
3.2.4.3 Removing stop words.....	32
3.2.4.4 Stemming	32
3.2.5 Feature Extraction Using Tf.....	35
3.2.6 Naïve Bayes Classification Algorithm.....	35
3.2.6.1 Gaussian Naive Bayes.....	37
3.2.6.2 Multinomial Naive Bayes.....	37
3.2.7 Split Dataset	38
3.2.8 Classification of Data	38
3.2.9 Performance Evaluation Methods	38

3.3 TOOLS	39
3.4 DEVELOPMENT ENVIRONMENT	40
3.5 SUMMARY	40
4. EXPERIMENTAL RESULTS AND DISCUSSION.....	41
4.1 INTRODUCTION.....	41
4.2 PERFORMANCE EVALUATION OF THE PREPROCESSING.....	42
4.3 THE CLASSIFIER MODULE.....	43
4.4 THE LEARNING MODULE.....	43
4.5 THE CORPUS.....	43
4.6 EXPERIMENTS AND RESULTS	44
4.7 EXPERIMENTS USING AN EVALUATION SET	48
4.8 SUMMARY	48
5. CONCLUSION AND FUTURE WORK.....	49
5.1 INTRODUCTION.....	49
5.2 RESEARCH CONCLUSION.....	49
5.3 RECOMMENDATIONS FOR FUTURE WORK	50
REFERENCES.....	51

LIST OF TABLES

	<u>Pages</u>
Table 3.1: word tokenization.....	30
Table 3.2: Example on word tokenization.....	30
Table 3.3: Documents after processing by normalization.....	31
Table 3.4: Example on Removing of Stop Word.....	32
Table 3.5: output stemmed.....	33
Table 3.6: Example on Word Stemming	34
Table 3.7: Example on Removing of Repeated Characters.	35
Table 3.8: Example on Removing Symbols.....	35
Table 3.9: Metric equation.....	39
Table 4.1: Documents used in the study.....	43
Table 4.2: Previous table describes each category.....	44
Table 4.3: Parameters used for NB 2225 document.....	44
Table 4.4: Classification accuracy on the evaluation set using TF with 2225 document.....	45
Table 4.5: Comparison of Results in the first trial.....	45
Table 4.6: Comparison of Results in the second trial.....	46

LIST OF FIGURES

	<u>Pages</u>
Figure 2.1: the sentiment analysis architecture[29]	14
Figure 2.2: Architecture of a generic sentiment analysis system[31].....	15
Figure 2.3: Feature Extraction with TF [36].....	19
Figure.2.4: supervised and non-supervised.....	20
Figure.2.5: Architecture of the proposed framework [42].....	22
Figure.2.6: Univariate Gaussian class-conditional distributions for two feature(heightand weight) and two classes (male and female) [47]	24
Figure.2.7: simple gaussian naïve bayes classification [49].....	25
Figure 2.8: simple multinomial naive bayes classification:.....	25
Figure 3.1: Sentiment Analysis System proposed (model).....	28
Figure 3.2: Overview of Data pre-processing.....	29
Figure 3.3: Word Tokenization	29
Figure 3.4: Word Stemming	32
Figure 3.5: Environment of Pc	39
Figure 4.1: The code of training and testing in the first trial.....	45
Figure 4.2: Training=60 and testing=40 of our model accuracy	46
Figure 4.3: The code of training and testing in the second trial.....	46
Figure 4.4: Training =70 and testing=30 of our model accuracy	47

LIST OF ABBREVIATIONS

NB	:	Naïve Bayes
BBC	:	British Broadcasting Corporation
TM	:	Text Mining
TF	:	Term Frequency
ML	:	Machine Learning
DS	:	Data Set
gNB	:	Gaussian Naive Bayes
mNB	:	Multinomial Naive Bayes
CD	:	Classification Document
SA	:	Sentiment Analysis
DM	:	Data Mining

1. INTRODUCTION

1.1. GENERAL VIEW

Firstly, we know that the categorization of the context of the text in documents is playing an important role in the range of data mining to reach absolute knowledge. It is relate with the way of organizing the documents, in addition to analyzing the texts it intrudes to show the ideas probably of the import of the context of text and the subject to which it belongs, in addition, to offer another topic that related to the main topic of the text , consequently it will contribute to set a suit title to the text that untitled text . the outcome of the massive development of technology through the world of knowledge, therefore, the indefinite amount of texts, documents, data, and articles or maybe personals works and literature are as tools of social communication, it will be showed prompt work to category the documents in order to the specialist's analysis the texts easy and simple [1].

Absolutely the document categorization contains two main folds, the first fold is defined by rebased categorization to access easily like any access needs to database that use as the main information to classify unknown texts and this type of categorization makes human experience to be able to create the discreet knowledge base. Another fold of document classification known by the teaching field by this kind, the group of known and labelled documents are used to teaching the machine what relates to each specific category so the phase is known as the training stage where the machine will be derived Frome the titled data and documents[2].

All views and opinion in general whether it is private or public about a big difference in several topics that are expressing via numerous social media, a lot of people even maybe almost peopleware using social media to express about their views and opinion in addition to their emotion and feelings that happen in daily lives. However, many people post anything on social media like social activities every day as their emotions and opinion by using applications and sites that provide an interactive forum when people effect on each other on another hand, social media consider a big chance to business as a platform to make touch with customers by social media to connecting with their perspective of products and service [3].

Analysing feelings or sentiment analysis, seeking opinion are that an arithmetic study of people's views, emotions, evaluation and attitudes with society and individuals, issues, topics and events and their characteristics using retrieval and remembering information and linguistics, it is used to focus on positive, negative or neutral opinions, emotions, and assessments. The task is

technically difficult and very useful in terms of science and practice. For example, organizations and companies always want to search for ideas or opinions of the target group or the consumer about what it produces and what it provides. Potential customers also want to understand the opinions of existing users before using a service or purchasing a product.

In fact, it extends from the mathematical process to the curricula and sciences concerned with the administration in addition to social sciences such as marketing, health sciences, cultural and political sciences, communication, finance, and even historical records, and for its importance to society as a whole. This spread is due to the fact that ideas, views, and opinions are the basis of most human actions and are key factors in our behaviour. We can define all of our thoughts and beliefs about reality, to make the choices we will make, to a large extent how others see and value the planet. For that obvious reason, whenever we want to make a decision about something that we like, we often consider other people's opinions and interests. This thing depends not only on the level of individuals but also on the level of institutions and organizations in many cases.

Now simply if people want to shop and buy a product or a commodity for that to know the quality of this product and its value, it has become easier to know by evaluating people and their comments to users on social media, moreover, searching for opinion sites and viewing them on the Internet and spreading the knowledge contained therein remains a challenge due to the distribution of different sites. Each site contains, as a general rule, a massive number of documents and texts, and it will be difficult for the user to identify relevant sites, as well as extract and analyse ideas and opinions in them. Therefore, automated mood analysis systems are needed, so there are many emerging institutions and organizations that specialize in providing mood summary services. Many large institutions have also created their own internal capabilities [4]. These practical applications and industrial concerns served as a powerful catalyst for research in the field of mood analysis.

Current research has provided many methods for various mood interpretation tasks, which include both controlled and uncontrolled methods. In a controlled environment, early studies used all kinds of controlled methods of machine learning (such as neural networks, reference vector machines [SVM], maximum entropy, naive Bayesian method, etc.).

Examples of uncontrolled methods are the different methods that use the vocabulary of mood, grammar analysis, and grammatical patterns [4].

A lot of researchers used several different designations for classifying patterns, like personal analysis, emotional analysis, opinion analysis, opinion analysis, and opinion extraction [4]. Sentiment analysis has an excellent role in areas such as teaching, e-commerce, and surveys [5]. organization and companies designate customer reviews and monitor social networks in order to make an opinion. And the decision about their services and products and take the necessary and appropriate measures at the right time and place. Many researchers said that the value of the shares, and thus the mood analysis in social networks were correlated, and the future price of participation is often predicted using feelings from microblogging, for instance, Twitter [6]. The sentiment lexicon includes a couple of sentences, phrases, and words showing negative or positive fields and mood. However, the use of the vocabulary of feelings for the categorizing of moods is not enough due to the different orientations of phrases in several ranges and fields. Further information on the problems associated with the use of the vocabulary of feelings was discussed by Liu [5]. Many machine learning methods are used to classify moods (the so-called corpus approach). Studies have shown that machine learning was better than vocabulary-based techniques, due to its flexibility of use, ease of processing, and high accuracy. in the issue of operation depend on machine teaching, the text is initially evaluated as a bag of words, then it is transformed into many functions. After this, machine learning processes, such as Naive Bayes (NB), are used for classification. The mood analysis has three layers: the document layer, the phrase layer and the aspect layer [5]. This study specializes in document level to improve categories and classify it in mood analysis. NB can be making its control by a learning algorithm that states that all functions are independent and equally relevant. It is widely used to classify documents.

Two aspects of models are used to categorize text without generality, that is, the polynomial and multivariate Bernoulli model. Despite the functions were not independent, the NB algorithm isostasy effectively used to solve various machine teaching problems compared to several complicated machine learning algorithms, NB was considered a simple algorithm that supported probabilistic theory [7].

1.2 RESEARCH BASICS

As synonyms, the concepts of “analysis of opinions”, “analysis of moods” and “analysis of subjectivity” are widely used; however, their origin is not precisely itself, and several writers believe while all connotations have various notions.

it is substantial to know several connotations in order to sense the difficulties encountered in the course of this work.

1.2.1 Definition of Main Concepts

A point of view can easily know despite a positive or negative opinion, opinion, position, affection or assessment of the state of an object (produce, place, individual, happening, company or subject) or part of this object of an employer or set of the employer [8] [9].

1.2.2 Tasks

Many missions are connected with the analysis of feelings. some of those are carefully attached and difficult to dive with, probably in order to have several common portions [6]. The most serious is mood categorization: also called mood direction, point of view direction, mood direction or mood contradiction [10].

1.2.3 Techniques

So many of the few works have tried to present different ways of practice mood. many of the collection is working on a review of several programs/missions that can be found in SA. Other works seem to be based on most CA topics.

Thus, all Feldman groups work in five main groups: analysis of sentiment at the document layer, sentiment analysis at the sentence layer, sentiment-based analysis of aspects, comparative sentiment analysis and acquisition of vocabulary sentiment.

1.2.4 Natural Language Processing and Information Retrieval in Sentiment Analysis

confering to Cumbria, mood analysis is usually seen as a so limited NLP problem, where you need to understand only positive or negative moods with respect to all sentence and base or subject. [11]. The recording level considers the text as a representation of the nature or complexion of it. This layer is called the mood classification function at the document level. [12]. functions and generalizing opinions Another example is the work presented in the article, which uses the popular controlled process of synthetic restless networks and reference machines

vector to classify the status, which was the result of emotions that are used thousands of times when searching for a fragment of specific information[13] .

1.2.5 Applications of Text Categorization

There are many applications that play an important role on this scale, some of that implementation is known as follow up:

1. The establishment of records: a large digit of resources of electronic articles has to lead to the requirement to urgently draw up and deal with a document in the suitable classification regular to ease the employee their easy use. It is believed in connection with the more leading documents for the classification of the text of documents [13], [14].
2. Classification of opinions. With the increase in the number of active users online in public and on reaction networks from each scale of the nature of assemble [13].
3. Categorization of the basic meaning of the word. Document classification plays an important role in uncovering the hidden meaning of text documents. Thus, it gives users a quick, as in a competition text or context that it has[15].
4. Filtering and classification of spam data: these types of applications are essential to simplifying the task. Users receive spam messages that may be fraudulent or sent by hackers [16].

1.3 RESEARCH PROBLEM

In the linguistics approach, sentiment analysis is taken into account to be categorization trouble. It involves natural language processing (NLP) on many levels, and in Hermits its challenges. There exists a good sort of application that would interest from its results, such as news analytics, marketing, questions, knowledge bases, and then on . Generally, opinions are often inspected from two perspectives: opinion holder, and Opinion reader. The scope of sentiment analysis in reviews is that the opinion of consumers who write an online review. The focus of this dissertation work is to conduct sentiment analysis on BBC, document by identifying classify documents to business, entertainment, politics, and tech. pre-treatment should be done to reduce the size of the training data set, which in turn may speed up the training process.

- Difficulty in reading texts and classifying them through computerized programs, or classifying them according to the nature of the feelings they contain, and it may take a long time to do it.
- It is not possible to compile text documents with an understanding of the nature of feelings that they contain in record time or through computer programs or electronic applications.

- There is no calculated system or electronic program that can analyse sentimental feelings in text writing, and works to classify them according to distinct and similar groups in record time.

1.4 RESEARCH OBJECTIVES

The essential goal of this study is to style and improve an efficient and site document automatic document categorization system that classifies site documents. the precise research objectives are:

- To spot an efficient and automatic text arrangement using Naive Bayes's way .
- To validate and evaluate the suggested an efficient and automatic document categorization system using Naive Bayes Method.
- The essential goal of this thesis is to make and create the selected system for classification.

Texts of a particular field in the BBC, this method category

The BBC document four categories of classifying the domains of business, entertainment, politics, sports, and technology, because of a specific problem and its reason it includes a set of particular properties from data mining using crude rough mathematics with a lexical approach to facilitate the selection of fractures and, thus, Thus, to develop the implementation of the record classification system.

Six steps were used to analyse moods, that is, pre-processing, identifying attributes, sharing data, and ultimately classifying.

1.5 THE RELATED WORKS

During this period, the volume of information exchange through social networks has increased, and most users openly shared their thoughts and pieces of information. This knowledge could be for the analyst or researcher a gold miner to extract valuable information for strategic choice and decisions. So there tons of interest has been created through the domain of feeling analysis (sentiment analysis), Some factors substantiated by study still date, that push the event of the research region, include, augmenting of machine teaching ways in normal language processing and knowledge retrieval, increase in World Wide site to show training datasets for machine teaching algorithms and therefore the inquiry of economic and intelligent applications that the world shows.

There is an outsized number of research tries that specialise in sentiment categorization, with different goals and different supervised and unsupervised ways. There are usually supported machine teaching moods.

- ShoushanLi, Chen, Y., Huang, & Zhou, G. (2010) [17]. It essentially depends on the automated sentiment analysis method for categorizing tweets. In this work, he added sentiment categorization is to classify a document consistent with the emotional polarities it is In linguistics, sentiment analysis is taken into account to be a classification troubles It involves natural language processing (NLP) on many levels, and in Hermits its challenges. There exists a good sort of application that would interest from its results, like news analytics, marketing, question answering, knowledge Bases then on. Generally, opinions are often inspected from two perspectives: opinion holder, and Opinion reader. The range of sentiment analysis in reviews is that the views of consumers who write online reviews. The focus of this dissertation work is to conduct sentiment analysis on BBC, document by identifying classify text to business, entertainment, politics, and tech. pre-treatment should be done to decrease the size of the training database set, which in turn may speed up the training method. The next important phase to do in order to train a model is to choose the feature. The features choice allows the creation of a set of unique conditions (features) during the group by excluding vague conditions. After performing the categorization model and testing it, the accuracy, and remembering parameters should also be estimated as computation tense. Moreover, a comparing of the applied algorithms should be shown according to the classification results.

Document classification relates to the task of “developing a ready-to-use system for automatically categorizing a text document into various categories associated with documents”. Thanks to the widespread use of the planet Wide site, large amounts of data on the Internet make up an excellent resource. The lack of organizing web documents makes retrieving relevant information from the internet an intimidating, time-consuming task and stimulate an event with automatic web document classification systems. Automatic document classification is a dynamic and difficult space to seeking, and once we assumed the algorithm, the Naive Bayes method (NB) in it. Favourable or unfavourable is a good instance for it. Sentiment categorization achieved state-of-art performance. vocabulary level, word and sentence level, phrase level and text level were the four different levels during which sentiment categorization was performed ,Timing calculation approaches (based on vocabulary) and machine learning approaches (based on corpus) were the two kinds of atrial level within the emotion classification.

- R Moraes, JOF Valiati, WPGO Neto in (2012) [16] . Proposed for automating the mission of classifying one subject textual review, document-level sentiment classification is employed for showing a positive or negative feeling. Generally, supervised ways have two phases:

1 .extraction/selection of informative features.

2 .Browse categories using educational methods such as Support Vector Machines (SVMs) and Naive Bayesian (NB).

SVM has been widely used and successfully as an approach to teaching emotions while artificial neural networks (ANN) have rarely been considered in comparative studies within the literature of sentiment analysis. This paper provided an empirical comparing between SVM and ANN in relationship to document level morale analysis. Experiments indicated that ANN produces superior results or a minimum of similar results to SVM.

- Singh, V.K, Biryani, R., Uddin, A, Waila in (2013) [18] The authors assessed and compared the performance of the three ways, i.e. machine-dependent classifiers (Naïve Bayes and SVM), and the non-controlled semantic guidance approach (SO-PMI-IR), thus Senti WordNet's approach to morale in movie reviews. It has been found that the categorization accuracy of NB is marginally better than SVM and is on the verge of the SO-PMI-IR algorithm. SO-PMI-IR's algorithm has obtained superb levels of accuracy and seems to be the only option due to its unsupervised nature. The writers conclude that NB's performance usually resembles the superior performance that is familiar to SVM and is the lower-rated sense rating qam almualifun bitaqyim wamuqaranat At an equivalent year. Suchdev, R., Kotkar, P., Ravindran, R., & Swamy, S.[19] . They analysed people's sentiments in their tweets and it makes the businesses mention profits in their products by increasing the quality and therefore the recommendation given by the customer's tweets. Such as any post or tweet it has two sides, Positive and negative tweets were noted and further objects have proceeded consistent with those records. Through this paper, they need to discuss the sentiment summarization system which takes the texts and articles that need to be analysed as input and generates an in-depth document summarizing the views within the input documents.

- Łukasz Augustyniak , Piotr Szyma , Tomasz Kajdanowicz, and Włodzimierz Tuligłowicz in (2015) Proposed a replacement method called "iteration" that automatically assesses user feelings (opinions) from the Amazon Reviews data set. As an extension of this method, they developed a dictionary of words, calculating the probabilistic repetition of words in the text and evaluating the effect of the recorded polarization, dividing the elements contained in the text. They analysed the results obtained using the glossary of unigrams, bigrams and trigrams using lexical, controlled and uncontrolled approaches to machine learning, and compared 37 methods of machine learning methods (glossary, groups, with vocabulary predictions as controlled trainers and observers). judge the analysis. Amazon dataset. Here, the authors claim to be one of the first cross-cutting areas of moral analysis in the literature[21].

Muhammad Bakali and Abdel Moneim Shukr, 2014 [22] , proposed a classification system for Arabic texts for Arabic tweets, mixing pure SVM math support with naive Bayesian (NB). This method in combination is applied to a group of Arabic tweets belonging to 5 categories. Evaluation and comparison in dimension F1 using the NB and SVM workbooks in TC.

Shugufta Fatima and B. Shrinivasu, 2017 [23] , An automatic text classifier of the SVM support-vector system of support has been prepared. This method includes all the previous steps for symbolic manipulations, pausing words and stopping the application. Before most SVM classification algorithms, functions within the system are weighted using the term TF. This technology has reached an average accuracy of around 86.39%.

1.6 CONTRIBUTION

The tries of this research to the state-of-the-art are multi-faceted. This system is designed specifically for the classification of English articles in an honest and important field (politics, business, sports, technology, and entertainment). this system supported two main aspects which are NB and have an extraction. the next points show the foremost contribution during this thesis as follows:

1. Proposed document stemmer to get the basis of words, Where the stemming features a great role in unifying the form.
2. Proposed of automatic future extraction.
3. Proposed of four important pre-processing.
4. Proposed of coaching of classifier supported modified NB.

1.7 THESIS ORGANIZATION

The last chapter is a general introduction to this thesis; it explains the points that organize the content of opposing chapters through this thesis.

- Chapter Two: The Title “Theoretical Background of Documents” Classification with the semantic universe document mining, document categorization and semantic approach with an explanation of all kinds, ways, and requirements of these methods.

-Chapter Three: Entitled “Development and implementation”, Articles denoting a system of semantic battles (Naive Bayes), offer the development and implementation of the proposed system.

-Chapter Four: “Experimental Results for a Naive Bayesian System” states the The results of the proposed system and explains the impact of evaluating system performance.

-Chapter Five: Entitled "The conclusion that future work" The conclusion is drawn from the final results of the thesis with a set of suggestions Future works will be suitable for this technology.



2. THEORETICAL BACKGROUND

2.1 INTRODUCTION

Working with an unstructured dataset, such as text documents inside the digital world, just like various websites and social media applications, includes this really important information, or statistically requesting a large number of opinions, ideas, or perhaps lifestyles that express who publishes it, represents a large target group For ordinary companies, even governments. Text categorization can be a process aimed at identifying specific, unspecified documents by categorizing them into the appropriate category to which you really belong. This process requires a lot of requirements to complete classification implementation, it starts with pre-word processing, feature extraction, feature selection, business training and ends with a test phase representing the ranking ranking level. This chapter closely explores all topics related to text extraction and document classification.

2.2 TEXT MINING

Text mining (TM) is one of the general significant methods that have lately emerged in handling deposited dataset, and this path characterize from data processing in which it deals with unstructured data, and this type of knowledge differs from structured data performed by the epidermal text, in other words, A purview is formed by a person to invert a specific motif or mention a particular theme and every individual comment and posts on Social media is a sample of unstructured data, while structured data is formed by specific dataset or unspecific, tatic, descriptive records dope about some of the files provided such as healthcare or commerce and other forms of databases, structured data have the only understanding of the machines because they include a set of qualifications and decisions, and this includes, it makes instrument from building real relationships between opposing information from unstructured data processing. Machines that encounter a problem affect this type of knowledge because they are made from detailed raw information, but people can easily recognize this dope, distinct machines. Typically, There are two main definitions or beliefs (TM)

The definition is to consider extracting text as a subfolder subject to data analysis, and this definition is setup on unstructured data in text analysis, which must be fixed or simplified text in order to create it, suitable for understanding the device. The second definition of text extraction describes the process of extracting text as a global file, and therefore, data processing is a secondary section of the text mining process in the primary view [24] .

2.3 BBC (BRITISH BROADCASTING CORPORATION)

The British Broadcasting Corporation (BBC) may be a British public service broadcaster, Formerly referred to as BBC is an online BBC service. This is a huge network of websites, containing well-known websites like BBC NEWS and Sports, joint video and broadcast services of BBC iPlayer, CBBC and CBeebies, as well as educational services such as Bitesize. The BBC has an online presence that only supports its radio and television programs and online initiatives since May 11, 1994, but it did not officially start until April 28, 1997, after the government agreed to fund it from revenue from TV licensing fees as a service in its title.

BBC Online includes a wealth of content, including news, sports, music, science and technology, and entertainment. The website has a British flair, although the homepage, news part and part provide different content between UK and “international” visitors. There are also detach pages for Wales, Scotland and Northern Ireland written by the BBC Nations.

The site concentrates on the first top-scale ranges: history, sports, weather, iPlayer, television and radio. Easily accessible from the taskbar, which is located at the top of the current BBC web pages.

Available only through the AZ index on the BBC website. These contain archives, art and culture, ethics, gardening, parenting, religion and travel news. However, there are more top-scale ranges — around 400 in March 2010, but this number has decreased since top-scale ranges are now often associated with a lower name which link to websites for specific services or programs. Since the emergence of social media within the press landscape, the BBC has endeavoured to provide reports more relevant to its audience while maintaining its authority as a public broadcaster in reporting crises. Using empirical analysis to produce crisis news on the BBC, this book shows that the emergence of social media on the BBC, and thus has acquired management of this type of material has led to alternative logic for the media during which smart journalists struggle for alternative central technology within the newsroom. In this changing context, political, economic, social and cultural logic led to a more relevant newsroom featuring this new category of journalists and BBC audiences. This examination of story production events explains that in the midst of shifts in journalistic practices and rules, including newsgathering, provision of sources, distribution, and impartiality, the BBC has reconfirmed its authority as a broadcaster [25] [26].

2.4 SENTIMENT ANALYSIS

Emotion Analysis also referred to as Opinion Mining is that the ongoing research within the text mining field. Emotion analysis comes under the category of textual mining which suggests mining the text in its simplest form in order that noise data are often removed and builds systems that attempt to identify and extract opinions within text. generally, there are two sorts of textual information is out there on the online. they are facts and opinions. Whatever we type within the search engines comes out as fact, which uses computer readable information like BBC tags like title, headings and metadata [27]. However, an equivalent cannot apply to opinions because opinions are being taken from the online that's available in several formats like reviews, blogs, news story, discussion groups then on. Sentiment analysis may be a combination of tongue processing, linguistics, informational retrieval, machine learning and AI.

However, an equivalent cannot apply to opinions because opinions are being taken from the online that's available in several formats like reviews, blogs, news story , argumentation groups then on. Emotion analysis may be a combination of tongue processing, linguistics , informational retrieval, machine learning and AI soon.

In order to have a successful or well-established business or event or event, it is important for companies or event organizers to understand the reaction and feelings of your target customers or people who have responded to them through social networks. Thanks to this world of technology, the expression of feelings, feelings and opinions about any situation has become much easier through social networks. The reaction of buyers and visitors on social networks is open and will contain reviews from them in the form of copies.

Thus, a general opinion regarding the popularity of work and fabric is quickly available on some social media blog. These blogs contain valuable information that will enable analysts to extract specific information through social media platforms. However, a typical process is needed to evaluate this achievement, and it is here that moods are mainly analysed. Moral analysis, along with Opinion Mining, are two processes that help classify and research customer behaviour and the approach to the brand, product, events, company and customer service. An analysis of moods is often determined through the automatic process of extracting emotions from a user's copy by processing unstructured information and preparing a form to extract knowledge from it. For the

time being, sentiment analysis may be a subject of great interest and development because it is so many practical applications[28]. Figure 2.1 shows the sentiment analysis architecture.

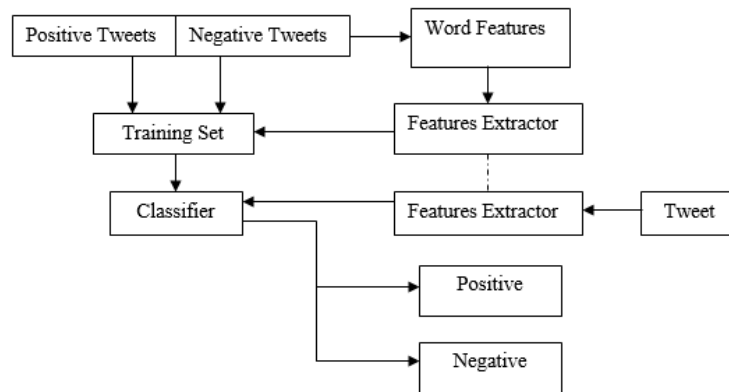


Figure 2.1: The sentiment analysis architecture [29].

2.4.1 Approaches for Sentiment Analysis

Emotion analysis indicates to the common process of extracting polarity and subjectivity from a semantic orientation that indicates the power of words and the polarity of text or phrases. There are two major ways to automatically extract emotions are dictionary-based and machine-learning approaches, Figure 2.2 shows the Approaches for Sentiment Analysis.

2.4.1.1 Lexicon-based approaches

The thesaurus-based process uses the thesaurus of opinion words and matches with the info to work out. They assign degrees of emotion to words of opinion that describe how positive, negative and objective the words in the thesaurus are. Thesaurus-based approaches are primarily believed in the glossary of meaning, i.e. a set of well- famous and previously translated codes, vacuity and expressions, advanced for classical types of connection, just like the Opinion Finder glossary[30].

2.4.1.2 Corpus-based

The corpus-based process aims to provide dictionaries for a specific area. These dictionaries are created from a set of preliminary opinion code that grow by searching for related words by using either statistical techniques or semantics [30]. Supported Statistics Methods: underlying semantic analysis (LSA).

Supporting semantic process, such as using alternative and opposites or relationships from synonyms such as WordNet, can be an interesting solution. Figure 2.2 shows the architecture of a general mood analysis system.

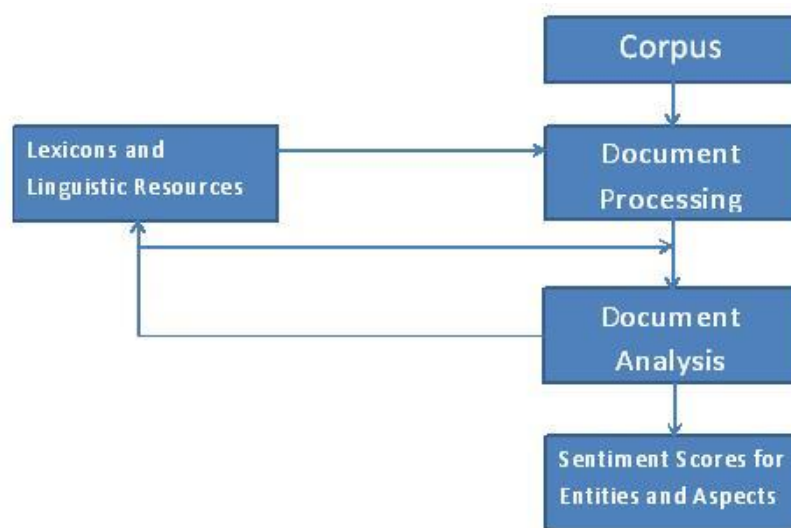


Figure 2.2: Architecture of a generic sentiment analysis system (from Feldman 2013) [31].

2.5 PREPROCESSING OF DOCUMENTS

Text pre-processing plays a critical part in (TC), which, before any method involving text analysis, filters out text words and removes all words that don't have the real point is to return to the topic of the document, in addition to this. it works to combine words formed by removing the word of any Language additions.

The purpose of the pre-processing step is to reformulate the text documents into the appropriate text coordination for automatic working and to increase the velocity and rigor, which leads to an increase in the execution of the text classification system. There are four basic implementation that represent the text pre-processing of tokenization, settlement, and removal of stop words that occur, which explains the following.

2.5.1 Tokenization

Tokenization is the method of removing spaces, numbers, pointing marks and tags from text and converting word to a schedule of components. This method effect a selection of words called characters. In some languages you like English and Arabic, words have transparent borders to distinguish one word from another through spaces [32].

The purpose of Tokenization i is to exchange confidential data - usually a payment card or check reckoning numbers - with a random number in the same format, but without its main importance.

This is different from encoding, where the variation is computationally variable, but the ethical template remains stored inside the new sign known as Preserve Format Encryption. Tokenization is the method of taking off sensible data from your working systems by changing it with unsaved code and saving the first data during secure storage for cloud data. Encrypted numbers are often decrypted using the received key. However, codes cannot be cancelled, since there is no arithmetical connection between the token and its ethical number.

Information is deleted in the invert order, replacing the token with the first number. Deletion is often overridden by the first token system. Nothing else thanks to the fact that I received the first token. Often, one-time codes (one-time open credit transaction) are not supported or are reusable (MasterCard number of a regular customer) and are stored in a database of repeated transactions. The purpose of the token podium is to get rid of any confidential payments or initial personal data from your business systems, replace each value with an unencrypted token, and save the first data in a secure cloud storage that violates your data environment. For example, after processing a payment using a token stored in your systems, only the first token system can switch the token with the corresponding PAN (primary account number) and send it to the payment processor for permission. Never register your systems, do not transfer or store PAN - only a token.

The token podium that comes with external data does not allow offense to receive any useful information - monetary or personal. "Useful dope" is this key. Although the token cannot guarantee the prevention of penetration, it can eliminate the sensitivity to data, which makes it useless for hackers. Token function and cloud storage data There is no information to be stolen in case of hacking. The risk of theft of knowledge is practically excluded.

2.5.2 Normalization

Normalization means the regulation of microarray data for effects arising from differences in technology, and not from biological variation between RNA samples or between printed studies [35]. This article describes normalization processes, confirming the reality that the equilibrium of dyes usually varies depending on the intensity of the spot and on the spatial position on the array. Loess tip-tip type normalization supply a fully-tested general aim normalization that gives good results on good array arrays. Tactics can also be improved using quality weights for individual locations. this tactic is best mutual with diagnostic information graphs that display spatial and intense trends. When diagnostic plots appear that show that offsets still remain in the data after normalization, further normalization phases can also be taken, such as normalizing the order of the plates or scaling the normalization between arrays. Compound normalization can also be used when there are control points that are known to have no differential expression.

Differences in loess normalization contain universal loess normalization and two-dimensional normalization. itemized orders are given for the implementation of normalization techniques using freely available software.

2.5.3 Stop Words

In every word that will be analysed, you can see that it includes several words that do not contain significant meanings, and there is no connection with most of the text, and these words are mutual in the language and are often repeated in the text.

Wording of words, such as words that connect phrases, as well as prepositions and conditions. All of the above words are unnecessary words, and you can, however, delete these words to decrease text size and preserve a set of serious functions that facilitate text sorting.

The method of inverting data to something a computer can understand is mentioned as pre-processing. one among the main sorts of pre-processing is to strainer useless data. In tongue processing, useless words (data), are mentioned as stop words[33].

2.5.4 Stemming

any word in the language has an ethical form called a root. to motivate the basis of each word to use one of the processes of definition, divide the word into parts from additions (prefixes and suffixes) and return the word to its root. for example, a player playing out of these different kinds of words is transformed [34].

The basis is that the method of decreasing a word to it is a stock of words that is attached to suffixes and prefixes or to the roots of words called a lemma. Stemming is vital for understanding natural language (NLU) and natural language processing (NLP).

Stemming can be part of lingual research in the field of morphology and the restoration and extraction of information AI (AI). Stemming and AI knowledge offprint meaningful dope from broad genesis such as big data or the online, as you may have to look for additional kinds of words related to atopy to get the simplest score. Stimming is a neighbourhood of requests and search incentive on the online.

2.6 FEATURE EXTRACTION WITH TF

TF which stance for Term Frequency may be a statistic procedure of evaluating the significance of a term in given documents. Term frequency (TF) refers to what percentage once a particular text surface during a record. The reverse record recurrence measures the word load within the document, i.e if the word is mutual or exceptional within the entire document. The TF intuition

follows that the terms that appear frequently during a document are smaller than terms that are rare. The TF uses the vector space Modelling technique for representing document text [35]. TF is employed in document classification, text summarization and recommender systems Among other things use cases . Figure 2.3 shows TF extraction function.

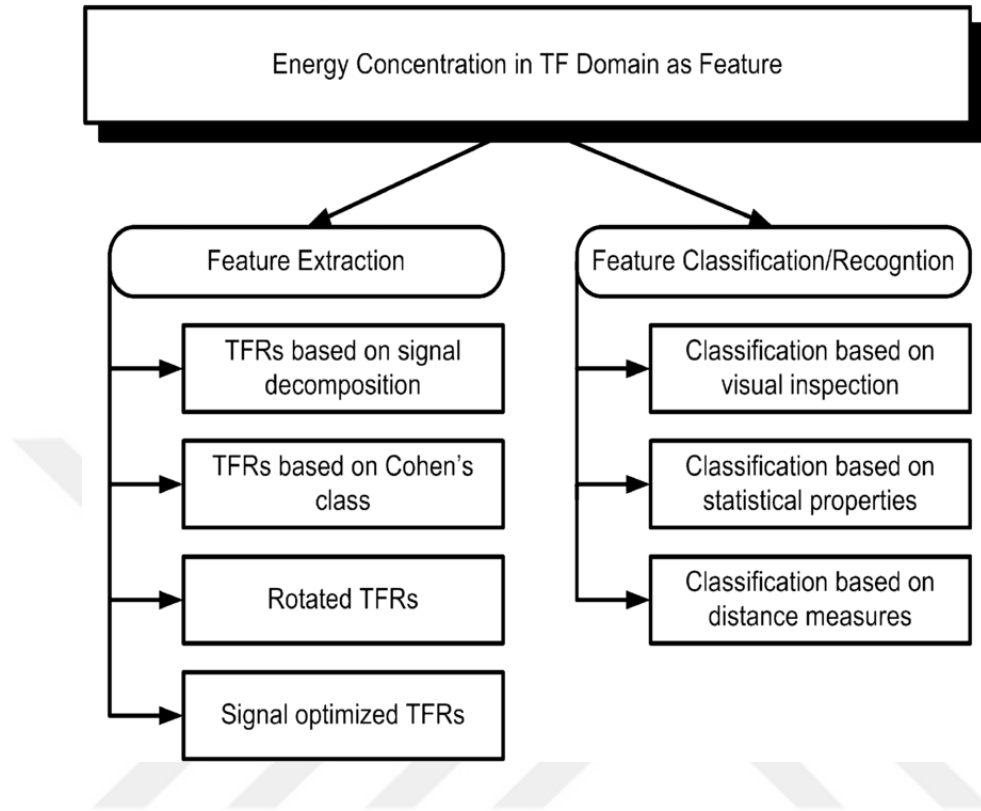


Figure 2.3: Feature Extraction with TF [36]

-Term Frequency (TF)

Term frequency (TF) of the term denoted as (t, d) is as follows: the total digit of times the indicated text t appears in record d , compared to the overall digit of all words in the record . If frequency of a given text increases during a record, then its TF also increases [36]. The content, format, and length of the text are different, these factors will affect the value of TF, and therefore normalization is the usual way to solve this problem. In practical applications, if a text function includes many stop words (the, an, my ...), and therefore, a high recurrence of the appearance of these states will increase the weight of the stop words TF, this will affect the classification. Results.

2.7 MACHINE LEARNING

Machine learning is an important approach in programming solution to form using of computers with highest performance and perfect efficiency depend upon prior knowledge or samples of

knowledge. The model in machine learning includes parameters and therefore the training phase or prior knowledge take the training task so as to enhance the parameters of the model [37]. The mathematical models in machine learning supported statistical theory essentially by using the samples to infer the results. The algorithms of the machine learning require two phase mainly: training phase and testing phase, within the first, be the implementation of the learning to extend the efficiency of certain algorithm and make it closer to the optimal. Second, after completed the training of the model, it's become able to provide solutions required, whether descriptive or predictive. The space and therefore the time could also be within the same level of serious with accuracy. Machine learning is divided into two main categories: supervision and unsupervised. Firstly, the training phase contains classify dataset of documents which contribute to support the testing phase, but in unsupervised, the documents classification execution with no prior knowledge. Decision tree, Support Vector Machines, K-Nearest Neighbour and Rough pure mathematics , of these methods that are mentioned considered a supervised machine learning techniques[38].

Despite the fact that the methods are offered under supervision and without censorship, the premier collection of documents/opinions, categorized, cannot be obtained. Moreover, hybrid approaches can combine supervised and non-supervised technologies, or even semi-supervised technologies. Can be used to categorize feelings.

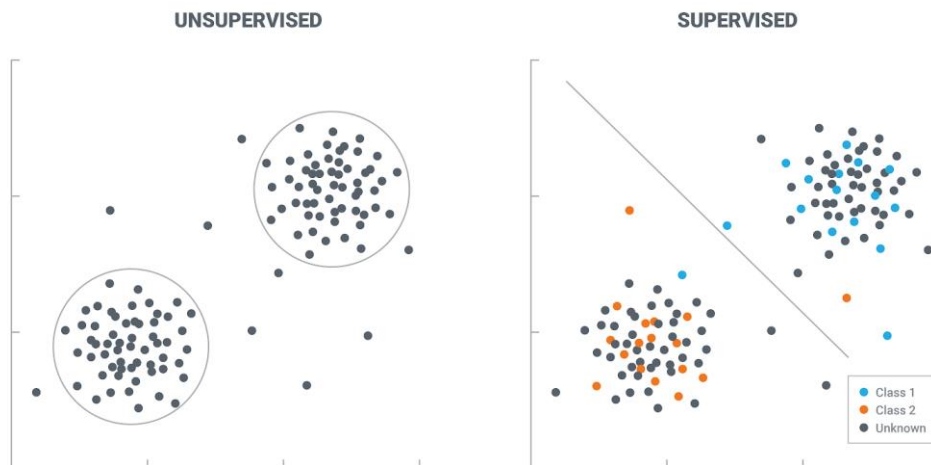


Figure 2.4: Supervised and non-supervised.

2.8 DOCUMENTS CATEGORIZATION

This procedure is often defined as appointing a collection of unnamed documents to the allocate category to which these documents reside. This is often the most common task of categorizing documents [39]. Categories are predefined; consider that specific terms differentiate each category from the others. the input document that matches the terminology of a category is then defined inside that category, and it becomes a tagged document. The classification of documents consists of two main parts for practice and workout:

2.8.1 Training Phase

Through classification operation, the machine learns the characteristics on which the divisions depend, so it is called the training stage. A collection of pre-classified documents is provided to identify features that distinguish each category. Input text objects are identifiers and are distributed in the form of vectors to include specific functions in the method in accordance with the algorithm used for classification, text functions are reduced, and the only thing that matters is It is saved by frequently highlighting text, so the process is performed using one of two methods for selecting a function. As an outcome, each class receives a unique collection of functions that simplified the classification operation [40].

2.8.2 Testing Phase

This is often the main classification process, since classified documents are classified and classified according to the accepted categories for each document. Classification depends on the characteristics of the training process. To evaluate the classification process, pre-checked documents must be pre-classified in order to match the results [2]. Before starting the classification test, text pre-processing should first be applied to documents that have been tested to standardize the characterization and compliance model.

Features obtained at the training stage to achieve high performance and effective classification [41].

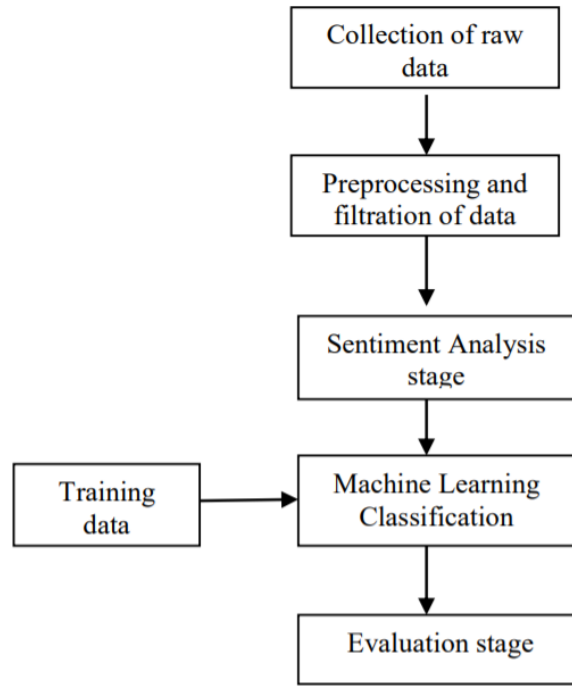


Figure 2.5: Architecture of the proposed framework [42].

2.9 DOCUMENTS CATEGORIZATION TYPES

Three of the most mutual way of classifying documents is Depending on the specific set of data that is explained below:

2.9.1 Hierarchal and Flat Categorization

Hierarchical classification and classification of apartments are two opposite types. A hierarchical approach that separates common classes, which includes a huge number of documents in subcategories, in more detail, including documents that have common characteristics, and this type facilitates the classification process. When using a massive dataset on the other hand, the flat classification method deals only with major groups and ignores all subcategories [43].

2.9.2 Multi-Label and Single-Label Categorization

A collection of data used, documents can largely belong to one category, and then use a book with several tags to detect this problem, so other previously identified documents belong to at least one category and are then marked with one label [44].

2.9.3 Hard and Soft Categorization

The motif of emotional rating includes techniques that you think are used in Measuring a certain significance to locate the relationship of any document to a certain class, but in a complex classification, the choice depends on whether the document belongs to the current layer at No, without significances for representing point of classification [45] [46].

2.10 NAÏVE BAYES CLASSIFICATION ALGORITHM

This research debate the opinion of naive Bayesian works and that application. Naive Bayesian collections are a set of complication algorithms that establish Bayesian theory. This is not a separate algorithm, but the algorithm family in which they all use the general standard, that is, each pair of classified features is separate from the others. Where d indicates the document that owns the classification, and c refers to the document category. The data set is split into two sections: a matrix of objects and a vector response. The basic premise of Naive Bayes is that each feature gives Equal, separate contribution to results.

As for our data set, this idea can be understood as:

We suppose that no pair of function is independent The purchase theory was mathematically shown as the following equation (2.1):

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (2.1)$$

$$P(y|X) = \frac{P(X|y) P(y)}{P(X)} \quad (2.2)$$

Now, it is time to make a naive supposition of Bayesian theory, which is a dependence between traits. To date, we divide the evidence into separate kinds.

Now, if each two premises A and B are dependence, then Therefore, we achieve the result:

$$P(y | x_1, \dots, x_n) = \frac{P(x_1 | y) P(x_2 | y) \dots P(x_n | y) P(y)}{P(x_1) P(x_2) \dots P(x_n)} \quad (2.3)$$

may be pronounced as:

$$P(y | x_1, \dots, x_n) = \frac{P(y) \prod_{I=1}^n P(x_I | y)}{P(x_1) P(x_2) \dots P(x_n)} \quad (2.4)$$

Now that the divisor remains stable for a specified input, we can remove this expression:

$$P(y | x_1, \dots, x_n) \propto P(y) \prod_{I=1}^n P(x_I | y) \quad (2.5)$$

In this study, we used two types of Naive Bayes algorithm, as shown in the following sections:

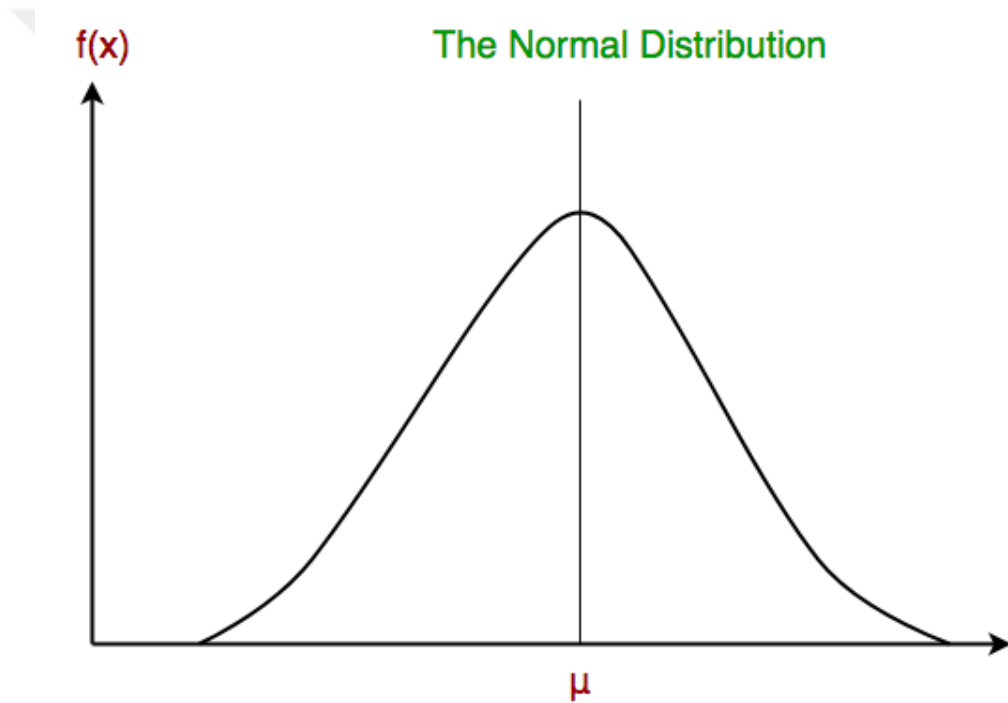


Figure 2.6: Univariate Gaussian class-conditional distributions for two feature(height and weight) and two classes (male and female) [47] .

2.10.1 Gaussian Naive Bayes

The Gaussian NB categorized, the meaning of digital characters is usually allocated. This distribution is presented relative to the average value and variance, which can help in calculating the prospect of the spotted values employ appreciation[48].

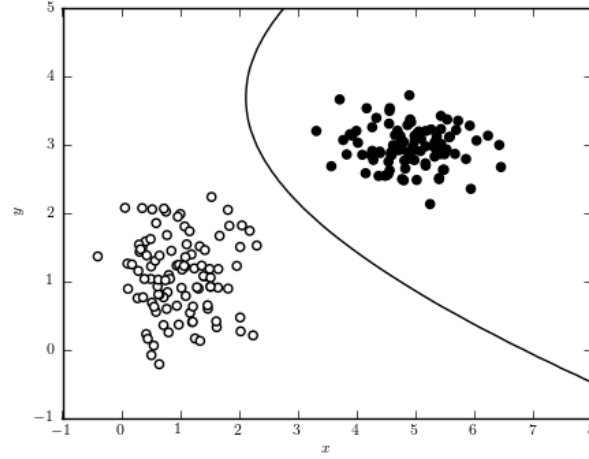


Figure 2.7: Simple gaussian naïve bayes classification [49].

2.10.2 Multinomial Naïve Bayes

A polynomial NB categorized for the indecision of the word knowledge is found in documents. It was installed according to a separate classification of functions. Since the polynomial distribution requires integer characters [50]

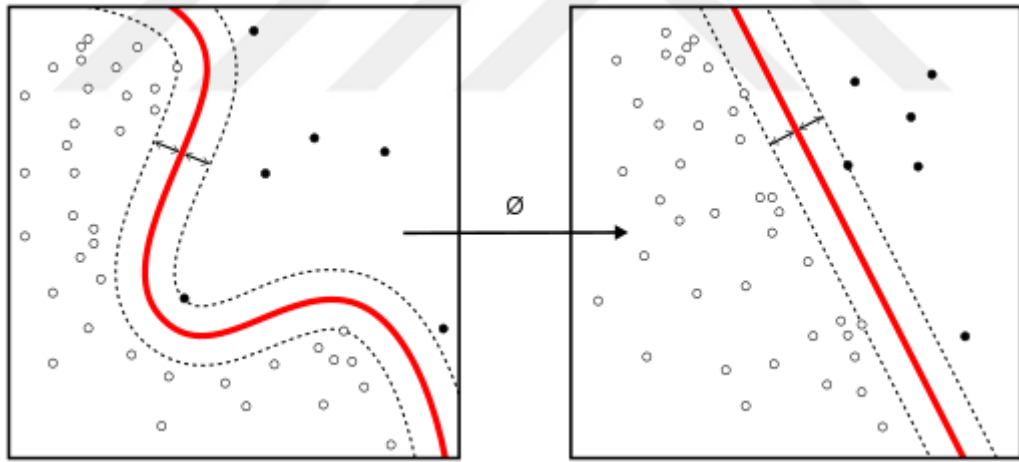


Figure 2.8: Simple multinomial naive bayes classification [51] [53]

2.11 SUMMARY OF THIS CHAPTER

In this chapter, we provided the background of our work associated with Twitter and movie reviews sentiment analysis. especially , we looked at how we will use sentiment analysis, social network and Twitter, pre-processing stages, feature extraction techniques, feature selection techniques. As for document extraction we will use TF-IDF Technique, and for the

choice technique we will use Chi Square technique. Finally, for sentiment analysis of data to business, entertainment, politics, sport, and technology, we will use NB.

3. THE PROPOSED SYSTEM

3.1 INTRODUCTION

Previous chapter provided an summary of the proposed methodology along with its design and implementation procedures roadmap to help the reader to quickly grasp the methodology.

This chapter extends the discussions by providing detailed and in-depth discussions of the proposed Categories document using Naive Bayes method. The remaining three sections are as follows. Section 3.2 underscores the detailed framework of the proposed research and Section 3.3tools. Section 3.4 development environment.

Section 3.5 summarises the entire chapter. Mood analysis, also referred to as opinion analysis, is an approach To Natural Language Processing (NLP). It involves the use of knowledge mining, machine learning (ML)[54] , , information retrieval and AI (AI) to mine text for subjective information and sentiment to identify business , entertainment, politics, sport and technology. neutral opinions, emotions and evaluations. This chapter describes the architecture and methodology of our mood analysis system.. This chapter contains the outline of all stages of the proposed system. Algorithms and techniques used at each stage will be explained. Each stage of the appliance of the sentiment analysis system include the required steps and algorithms where phase builds on the results of the preceding phase. The implementation details are going to be illustrated during this chapter.

The goals of study is the first to consider mood analysis in the BBC and document reviews to work out the attitude of the document is business, entertainment, politics, sport, and tech. towards the topic of interest, and second, is to develop and implement an algorithm for automatic classification of text into business, entertainment, politics, sport, and tech. using Naïve Bayes (NB).

3.2 THE PROPOSED SYSTEM

In this part, we briefly depict the complexity of the proposed system for analysing moods. As shown in Figure 3.1, the proposed system operates in two stages: training and testing stages. The purpose of the training stage is to create a classification model to distinguish between business, entertainment, politics, sports and technology. Documents supported the input of tagged documents by collections.

at the testing stage, the trained classification model will determine business, entertainment, politics, sports and technology. sticker on new unlabelled documents. The system contains five stages: pre-processing, feature extraction, classification model for analysing mood and, therefore, assessment of the system .

At the first stage: the very purpose of this step is to using Natural Language Processing (NLP) methods to process the text of the input document and make it suitable for the next steps for the correct extraction of functions. in the second step: there are several functions that must be extracted to represent the text of the input document. At this stage, we will present various types of functions that will be used in the proposed system using TF. in the third step: at this stage we use the chi-square as a method for selecting objects, its task is to select a subset of objects from the original set of functions that provide a constant value without losing information. in the fourth stage: the most important step in the framework of the proposed system is to create a classification model that is ready to effectively distinguish between business, entertainment, politics, sports and technology. flagged document at the training stage. There are several neural network algorithms that will be used to build such a model. as part of the proposed system, we have introduced a naive Bayesian (NB) classifier. These algorithms are widely used and have great success in text classification tasks. Finally, the fifth step: the system is evaluated using several performance indicators.

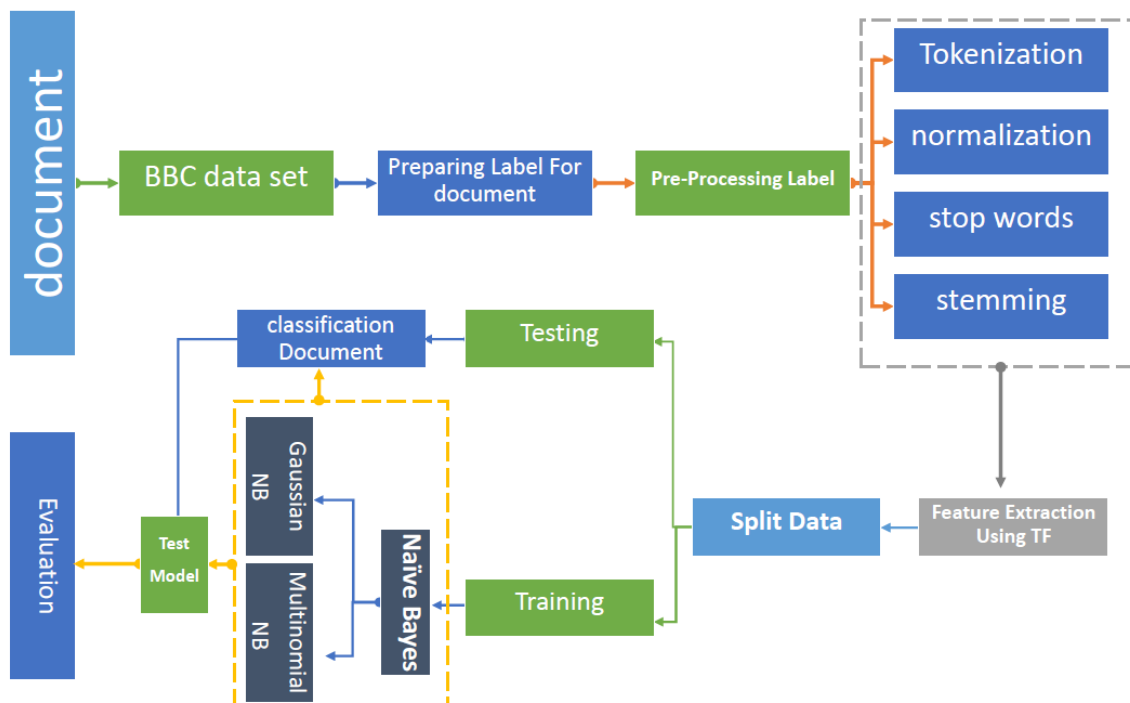


Figure 3.1: Sentiment analysis system proposed (model).

3.2.1 Downloading Data

I can use an open-source BBC data set, located here: <http://mlg.ucd.ie/datasets/bbc.html>. The data collection contains 2225 articles from the BBC News website. Each document is categorized into one of the following five chapters: politics, entertainment, business, technology, and sports. I downloaded the dataset and wrote the import-date script to output the data as two-column data Frame, with one row for each material. The first column contains the text of the document, while the second column contains categories.

3.2.2 Dataset Description

In this thesis, we classify sentiments with the assistance of machine learning. For the aim of coaching and testing of the system, different dataset corpus are used. Raw documents were gathered from the web site BBC <https://www.bbc.com/news>. Our data set comes with two tags for 2225 of 21162 documents, divided into NB_Mult and NB_Guz , and split this data to 40% training and 30% test dataset. The length factor of the feature was for the info set of document reviews.

3.2.3 Preparing Label

A label ticks a category or class of information. if the marking process is simply and easy to consider, it is stay an important section of data pre-processing in machine-learning algorithms . The training sets of our model are depended on original data with predetermined base characteristics (values). the chart is an important section of identifying targets using algorithms. The labelling operation must be carried out accurately, since any error or inaccuracy can lead to a standard dataset and, therefore, to the overall performance of the predictive model.

3.2.4 Pre-processing

Pre-processing is the first step of the proposed system. Text pre-processing is substantial stage in the sentiment analysis of the BBC documents' data. A Text may be a short message full of randomness like irrelevant symbols, misspelled words, stop words, and slang words. Also, number of these characteristics apply on movie reviews data. Those random characteristics considerably affect the performance of sentiment analysis approaches. So, to stop this, we apply pre-processing before feature extraction and have selection and use specific machine learning algorithm [33]. The most general view for pre-processing are often shown in Figure 3.2.

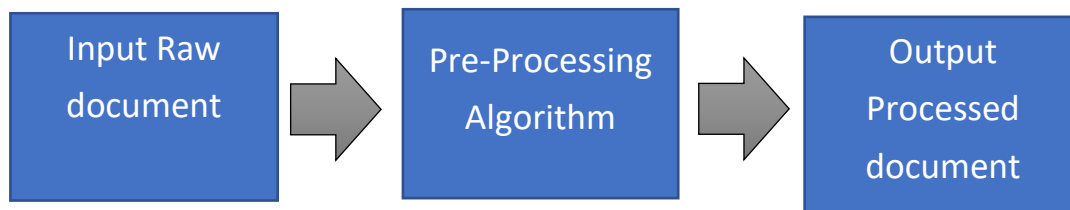


Figure 3.2: Overview of data pre-processing

After collecting my raw dataset, I processed the document text. pre-processing implemented using Python programming language and that i have applied number of pre-processing procedures to scale the data set and filter noise from the data and decrease its volume. We first do some pre-processing on document which is as follows:

3.2.4.1 Tokenization

This process is very important for Data pre-processing. In tokenization, a text document will be divided into tokens using a space to detail each term from others. Tokens can be numbers, characters, and terms.

Here, the tokenization package uses the Natural Language Toolkit (NLTK) tokenization package. The choice of a tokenizer depends on what knowledge you use, and therefore on the language. Here, the tokenization package uses the Natural Language Toolkit (NLTK) tokenization package. the choice for choosing a tokenizer depends on the characteristics of the information you are performing, and therefore on the language. Here, to make a tokenization method for tokenizing words using a document tokenizer module to process English terms. The word tokenization algorithm using the document tokenizer is shown in the algorithm below (3.1).

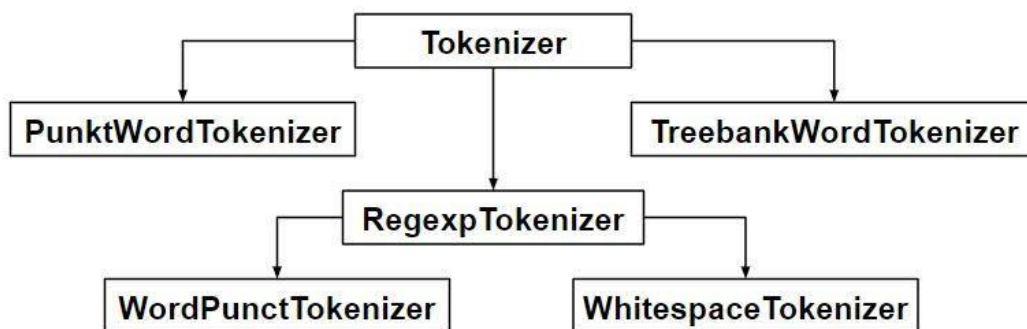


Figure 3.3 Word tokenization

Table 3.1: word tokenization.

Input: Documents
Output: Tokenize words
For all the words in the processing documents
Add the token status
Connect the word by going to Document Tokenizer Method and appending
Tokenize Sentence
Return a wholesale symbol

The tokenized text after removing irrelevant information from the original document is shown below in the table (3.2).

Table 3.2: Example on word tokenization.

no	Original Document	Word Tokenization
1	is so sad for my APL friends	['is', 'so', 'sad', 'for', 'my', 'APL', 'friends']
2	Fifa ticketing just times out. Seems like my 2010 ticket chances are getting less and less	['Fifa', 'ticketing', 'just', 'times', 'out', 'Seems', 'like', 'my', '2010', 'ticket', 'chances', 'are', 'getting', 'less', 'and', 'less']

3.2.4.2 Normalization

This step is used to normalize information in an array of separate changing and assign. This execution is generally performed over the pre-processing of information. Normalization includes three stages [55], starting with the elimination of diacritics from the word and finish with the transformation of the literal term to others. In table 3.3 shown that document after processing by normalization.

Table 3.3: Documents after processing by normalization

Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

3.2.4.3 Removing stop words

When processing natural language, a number of words that are often found in documents are stop words (for example, “on”, “the”, “am”, “are”), which have only a small emotional meaning and do not affect the assessment of opinions and moods in relation to lexical resources. In my research, a pre-defined list of quit terms is used to discover these terms and remove them from any documents. The list contains all possible stop words in English.

The roster also includes many class: conditional pronouns, interrogative pronoun s, prepositions, referral nouns, pronouns relative, the transformation of verbs, verb pronouns and others. Therefore, employ stop words removal in the mood analysis process can be a necessary step in any system . We get a result after removing stop from the document makes the result more accurate. The result obtained after the stop words are removed from the document is shown in the table (3.4).

Table 3.4: Example on removing of stop word.

no	Document Tokens without filtering stop words	Document Tokens with filtering Stop Words
1	['is', 'so', 'sad', 'for', 'my', 'APL', 'friend']	[<u>'sad'</u> , 'friends']
2	['Fifa', 'ticketing', 'just', 'times', 'out', 'Seems', 'like', 'my', '2010', 'ticket', 'chances', 'are', 'getting', 'less', 'and', 'less']	['Fifa', 'ticketing', 'Seems', 'like', <u>'ticket'</u> , 'chances', 'getting', 'less']

3.2.4.4 Stemming

Justification and lemmatization of words is a method that creates a limited type of word in a text. Subtraction can be a technique that gets the (main) root of a word in a text. This normalizes the word by removing the suffix from the word, which provides the root of the word. There are many commonly available word processing algorithms. During this knowledge pre-processing approach, the Porter Stemmer algorithm is used to remove the word suffix to get the correct meaning from the text.

To implement this kill function during this search, the Stemmer Porter algorithm provided in the Python NLTK package is used and develops a function that returns a word after extracting all the letters during the word. Porter Stemmer comes from a word, letter by letter, removes the suffix and

provides the basis for the word. Here, in the process of fumigation, the word will be stopped and the basic meaning of the word will be returned. To achieve accuracy in moral analysis, just stop the word whose length is more than two, because words such as “a”, “is”, “OH” are not taken into account in the dictionary of meanings to get the word polarity. The root word execution algorithm is presented in the algorithm below (3.2).

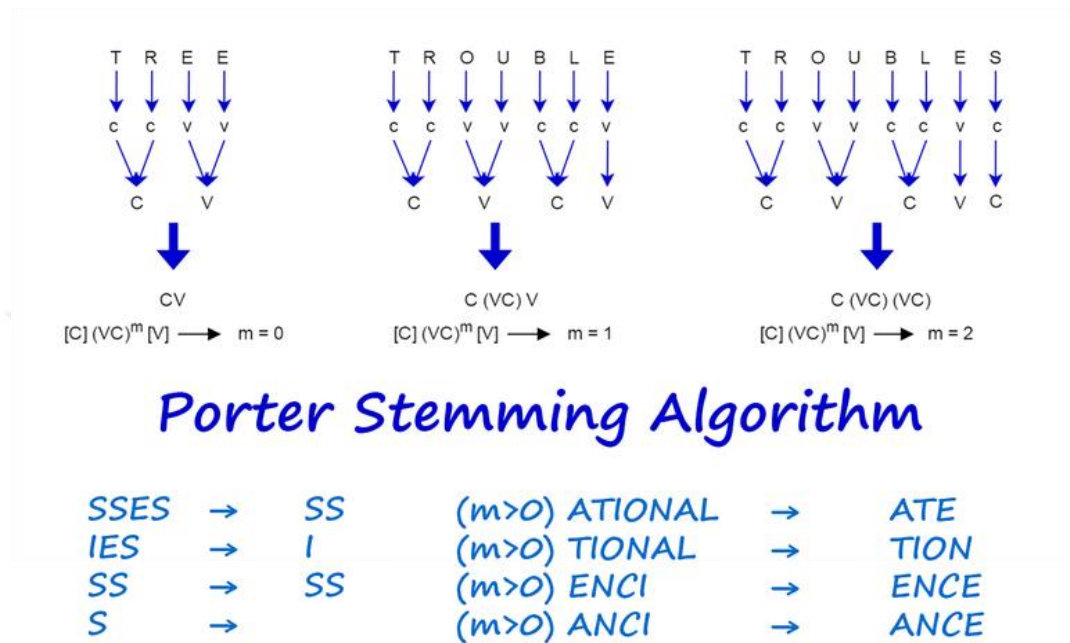


Figure 3.4: Word Stemming

Table 3.5: Output stemmed

Input: Tokenize words
Output: stemmed words
For word in word Tokens
Initialize Stemmed word variable to empty list
If length of word greater than 2
Method call for stemming the word using Porter Stemmer object.
Append Stemmed Sentence list
Return Stemmed word List

In this algorithm, the base of the word is longer than two, and add the word to the list of variable types to avoid returning one character. and the returned word in the form of a basis will be in an additional general form and can be applied in subsequent stages of the Natural Language Processing (NLP) task. In addition, the words “stemming” and “lemmatizing” give the general basic look of the word, removing the suffix from the word that provides the vocabulary targeting the word, for example, the words “cars” become “car”, shortening the letter “s”. In addition, words such as “consultant, consultant, consulting, consultant, consultant” are associated with the only word “consultant”, which is that it is a more general form of the word. this can give us a more general assessment of moods and help us evaluate the exact polarity of moods for a text document. The full result obtained using this algorithm is shown in table (3.6).

Table 3.6: Example on word stemming.

no	Document Tokenization Without Stemming	Document Tokenization With Stemming
1	['sad', 'friends']	['sad', 'friend']
2	['FIFA', 'ticketing', ' Seems', ' like', 'ticket', 'chances', 'getting', 'less']	['FIFA', 'ticket', ' Seem', ' like', 'ticket', 'chance', 'get ', 'less']

- Removing Symbols and Repeated Characters

We also found that there are some features that would affect my experiment' result, including “http://t.co/NxSE1HYTHj”, or (: \ | [] ; : { } - + () < > ? ! @ # % *,.). Many documents contain non-alphabetic symbols like “@” or “#” also as active web links. The word immediately following the “@” symbol indicates a username, which we filter entirely. The username is deemed to feature no sentiment value to the text.

in order to precise their strong feelings, people often use word with multiple characters. for instance, “LOVEEEEE”. the amount of ‘Es’ during this word are unnecessary and are therefore, required to be eliminated. The algorithm for removing document symbols and repeated characters is demonstrated in below algorithm (3.6).

Table 3.7: Example on Removing of Repeated Characters.

no	Input document without Removing of repeated characters	Output with Removing of repeated characters
1	my life is very lazzzzzyyyy!!!!!!	my life is very lazy
2	nothing to do alll dayyyyyyy.	nothing to do all day

In table (3.7) shown that document after removing of repeated characters.

Table 3.8: Example on Removing Symbols.

no	Input Tweet without Removing of symbols	Output with Removing of symbols
1	(must I say more?)	must I say more
2	A dog riding the bicycle http://bit.ly/gvMzD	A dog riding the bicycle

3.2.5 Feature Extraction Using TF

The term frequency (TF) refers to the frequency of repetition during a special record . Since the tallness of records changes from one to other, usually a certain term is refined most over in extended than in low records. To understand normalization, the code recurrence is generally split into a document using the next equation (3-1):

$$TF = \frac{n}{P} \times \frac{1}{T_n} \quad (3.1)$$

In which n appear the recurrence of expression in a clause, however, T_n appears the overall number of expressions in a record.

3.2.6 NAÏVE BAYES CLASSIFICATION ALGORITHM

The project applied a sample that conferring the provisional prospect substantive by in the equation (3.2).

$$P(c|d) = \frac{P(d|c)}{P(d)}, P(d) > 0$$
$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (3.2)$$

Where in d indicate to the record to be distributed, and c speak the category of the document. In practice, it was visible that $P(d|c)$ did not display much impact, therefore it was rejected. We advanced the Bayes judgment as substantive in the formula (3.3).

$$P(c|d) = p(d|c)P(c) \quad (3.3)$$

For each document, they calculated the meaning of the functions or statement in the record employs the equation substantive below:

$$P(c|d) = P(x_1, x_2, x_3, \dots, x_n|c)P(c)$$

$$P(x_1, x_2, x_3, \dots, x_n|c) = P(x_1|c).p(x_2|c).p(x_3|c) \dots P(x_n|c).p(c) \quad (3.4) \quad (3.5)$$

Where; d = functions x_1 to x_n in the document. To calculate each word during a wheel in a record, we applied a formula. (3.2). the NB classification, the categorized combined the suggested sample with the preferred decision rule. This judge has chosen the more likely thesis. The identical Bayesian classifier indicates the task to which the label category is assigned $y = c$, according to formula (4) below.

$$y = P(c_k) \prod_{i=1}^n P(x_i|c_k) \quad (3.6)$$

Where in y = final category famed for the trail that contrasted to the best touch of the info. We applied Eq. (4) for define, the utmost or floor y supported the NB applied. They also studied the account for $P(c_k)$ and $P(x_i|c_k)$ employs the two mentioned shapes as $P(c_k)$ was himself for all NB sort were within the account of $P(x_i|c_k)$ vary among the algorithms, confirming the classifier used.

$$P(c_k) = \frac{\text{document}P(c = c_k)}{N_{doc}}$$

$$P(x_i|c_k) = \frac{\text{count}(x_i|c_k)}{\sum_{x \in V} \text{count}(x|c_k)} \quad (3.7)$$

The main disadvantage of NB was that if there is no category label and the choice refer to the value, then the prospect value setup on the indecision will be 0. We applied the Laplace smoothing method for label 1 to solve this problem below.

$$P(x_i|c_k) = \frac{\text{count}(x_i|c_k) + 1}{\sum_{x \in V} \text{count}(x|c_k) + 1} \quad (3.8)$$

Two sorts of NB are utilized in this research as a report in subsequent parts.

3.2.6.1 Gaussian naive bays

In the Gaussian NB category, the value of numerical refer is usually deal. This allocation was offered in terms of medium and standard perseveration that will help in calculating the prospect of the spotted values using respect. The chance of brow was studied as:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right) \quad (3.9)$$

average and where μ = mean and σ^2 = variance was applied. persistent data is treating by discretizing the assigned rate using the sling process.

3.2.6.2 Multinomial naive bays

A multinomial NB category catches the frequency of knowledge words present in records. it was adapted to the separated rating of symptoms, since multinomial allocation requires integer functions.

$$P(x_i|c_k) = \frac{n!}{\prod_{i=1}^{|V|} x_i!} \prod_{i=1}^{|V|} P(x_i|c_k)^{x_i}$$

$$P(x_i|c_k) = \prod_{i=1}^{|V|} P(x_i|c_k)^{x_i} \quad (3.10)$$

Where: v = no functions. In shape, normalization $\frac{n!}{\prod_{i=1}^{|V|} x_i!}$ not rely on complication k . then $P(x_i|c_k)$, it is calculated using the equation. (3.7) after collect parameter V ., The Laplace equation was applied to stop a problem with 0 frequencies.

$$P(x_i|c_k) = \frac{\text{count}(x_i|c_k) + 1}{(\sum_{x \in V} \text{count}(x|c_k)) + |V|} \quad (3.11)$$

where V = dictionary in all training dataset classes.

3.2.7 Split Dataset

In my study, we divided our data into two groups: testing and training. The 70% split for training, while the 30% split for testing. And we did several other experiments that have value training and testing sets.

3.2.8 Classification of Data

There are several sorts of classifiers used for machine learning techniques to identify sentiment analysis of knowledge. Machine learning offers an answer to the sentiment classification problem. We used a method for classifying Naive Bayes based on polarity using a set of business, entertainment, politics, sport and technology documents for analysing moods. Polarity is determined by the ratio of the probability of occurrence of a word in a set of business , entertainment, politics, sport and technology.

Naive Bayes (NB) has developed as an essential tool for classification, for sorting a document during a neural network. therefore, categorization decisions are determined by the final values at the output nodes[56].

3.2.9 Performance Evaluation Methods

Statistical analysis is extremely important for measuring the performance of sentiment analysis systems. The accuracy of sentiment analysis represent the essential parameter to see the system performance. Moreover, and while the sentiment analysis is taken into account as a classification problem, the performance of the system might be evaluated in terms of the many parameters. during this thesis, four parameters are wont to check the performance of the proposed system This is resolution, precision, call, and F1 result. In Table 4, accuracy usually established using a confusion matrix. Here the matrix confusion depended to choose of data sets. In this research, we use an emergency schedule to improve accuracy and productivity.

Table 3.9: Metric equation

Metric name	Equation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
F1-score	$\frac{Recall + P}{2 * Recall * P}$

3.3 TOOLS

The program was in written the Python because Python can be powerful programming. The info structure and object-oriented programming concepts help the programmer find an efficient and effective way of programming with a minimum number of lines of code. Since it is open source, its interpreter allowed you to directly code your program, and also allowed us to access several standard libraries and resources that are freely available online to meet your application development requirements.

NumPy library is a necessary package for learned calculate in Python , which is written in Python and is therefore easy to use within Python [57]. [58]. It is also the most suitable language for writing scripts and developing applications, taking into account its complex syntax and dynamic typing, which is more pleasant for processing linguistic data [58], . During this study, Python 3.7 was used to process linguistic data using the NBTk (Natural Language Toolkit), which is most interesting in this version of Python.[59], Another library used is the Matplotlib, is a 2D plotting library for the Python programming language, Pandas is a software library written for the Python programming language.

3.4 DEVELOPMENT ENVIRONMENT

In this chapter we provide the results obtained by conducting the experiments we start by discussing the general. To test the suggested process, we formed an installation with the following system requirements. We tested our approach on Windows platforms.

System	
Processor:	Intel(R) Core(TM) i5-5300U CPU @ 2.30GHz 2.30 GHz
Installed memory (RAM):	8.00 GB (7.88 GB usable)
System type:	64-bit Operating System, x64-based processor
Pen and Touch:	Touch Support with 10 Touch Points

Figure 3.3: Environment of Pc

3.5 SUMMARY

This chapter rendered an in-depth description of every step of the proposed methodology for accomplishing the stated objectives. After a general introduction it underscores the detail framework of the proposed research, followed by during this chapter we have a careful description of the methodology that we have used in this thesis. we have collected a corpus of business, entertainment, politics, sport and tech documents from site BBC .Then we apply the pre-processing stages. the most The goal of this step is to use many processing stages, the pre-processing step includes sub-steps like (tokenization, Abbreviation, Removing Stop words, Stemming) To process the input document text and make it appropriate for the next step to properly extract features. After pre-processing stage comes the extraction feature stage. The Feature of the extraction is that the convert of the exporter facts into a knowing Sirius with a diminutive number of changes, which primarily contains the special reference. this can lessen the dimension of information, stripe excessive or irrelevant data and convert it into a form more suitable for subsequent classification. during this dissertation, we use the TF process to excerption the function. Finally, we first apply Naive Bayes (NB) to our training kit, then trail and contrast the effects.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 INTRODUCTION

Classification of categories of documents is that the process of summation documents in the classification depends on their content. document Categories characterization may be a huge learning issue that is at the centre of various data the board and recovery errands. Document Categories plays out a basic job in several applications that manages checking out, characterizing, looking, and succinctly speaking to many data. Document Categories may be a longstanding issue in data recovery, which has been very much contemplated.

This chapter attends to separate the empirical effects, permission, and dispute of the performance of purposes. The effectiveness of the offer “Categories” record using the naive Bayesian criterion is evaluated using several sizes, as confirm in chapter 3. in this respect, a string of experiments are managed to employ overtly ready category document datasets using naive Bayesian. The qualitative measurements are supported visual human inspections, while the quantitative measurements are performed using standard measurement approaches. additionally, Categories document using Naive Bayes is benchmarked against the simplest and up-to-date techniques found within the literature. The chapter consists of four broad sections and organised as follows. Section 4.2 discusses and presents the automatic classification of Categories document using Naive Bayes. The pre-processing of Categories document using Naive Bayes in underscored in Section (4.8) concludes the chapter.

This section represents the results gained with several regulation from the foremost major method to affect the foremost developed applied during this study. Therefore, the part likewise validates the need for the examinations speak around and the way they assist in developing the approach. The sample for document characterization is testing versus a experience group of statement. The fill of the sample was chosen using the measurements shown next:

$$Accuracy = (1 - \mu / N) * 100\% \quad (4.1)$$

Where μ is different, it is wrong to request a record from a test set control N archives. all result indicates a separate classifier run. The sample is complied with tested for several sizes, such as Precision and Recall. Accuracy (P) is often term due to the number of true positive results (T_p)

compared to the number of false positives (Fp) in addition to the number of true positive results (Tp).

$$P = \frac{T_p}{T_p + F_p} \quad (4.2)$$

Recall (R) realizes the digit of True Positive (Tp) through the digit of False Negative (Fn) excessive the digit of True Positive(Tp).

$$R = \frac{T_p}{T_p + F_n} \quad (4.3)$$

These values are also associated with the estimate (F1), which is defined as the harmonic mean of accuracy and recall.

$$F1 = 2 \frac{P \times R}{P + R} \quad (4.4)$$

All indicators obtained vary from 0 to 1, where 1 is ideal and 0 is the worst.

4.2 PERFORMANCE EVALUATION OF THE PREPROCESSING

pre-processing. As a primary step in pre-processing, authors deleted every HTML label for the online record. beyond that, the writer remote whole stop words from the record . Stop words are common terms, such as “a,” “a,” “to,” etc., which do not support in the classification, as they show in every document in any case of the class to whose they stun. The authors after deleting the word stop the execution of the trunk. After that, we selected the 20 most frequently occurring words during the document to present it.

The steps for pre-processing and classifying a replacement document often come down to the following:

- Remove dots, virgule, pointing, stop words. assemble the terms that show within the record in all once. We purport this set of terms a dictionary.
- Consider often occurring words as a selection of terms, coordination terms detected in the dictionary as well as learning set documents.

- Search for the corresponding set (s) of words or its subset (control full whit) in the list of sets of words compose from training data for which in the subset (points containing only one) of the

set of frequent words of the new record.

- Collect the appropriate chance meaning for the kit (s) of matching words for all goal sort.
- Calculate the chance sense for each aim grade from the naive Bayes class proposition.

4.3 THE CLASSIFIER MODULE

The simplistic Bayes method (NB) may be a simple Bayesian workbook that supports Bayes's theory of applied mathematics. Within the categories document using the Naive Bayes system, root types of words that are used in training documents because of the characteristics that each document represents.

4.4 THE LEARNING MODULE

The primary purpose of the training module is to pass tags with ordained categories from the record collection to allow the classifier to classify modern random D records and assign them to all the specific target classes C_i . This meter is configured based on the NB learning algorithm definite.

4.5 THE CORPUS

For this study (2225, 21162) the attachment was used. data collected from various sources such as the BBC documents. There are three labels in this dataset which covered topics like politics, business, sport, technology, and entertainment. The table 4.1 describes this dataset.

Table 4.1 Documents t used in the study.

Category	Document number
Business	510
Entertainment	386
Political	417
Sport	511
Technology	511

At the following Table 4.2 i will present the number of the documents the i uses in every category.

Table 4.2 Previous table describes each category.

No	Category	Text
1	Business	WorldCom boss left books alone former world C...
2	Sport	tigers wary of Farrell gamble Leicester say ...
3	Technology	tv future in the hands of viewers with home the
4	Politics	Stewart white with the latest Politics news and debate...
5	Entertainment	ocean s twelve raids box office ocean s twelve...

4.6 EXPERIMENTS AND RESULTS

Simulations help in the study of the application of machine learning algorithms. There are several parameters that are used with NB, as shown in Table 4.3. We used the confusion matrix in order to determine accuracy. There are experiments that are performed during this work. TF is used with two NBs. TF applied and obtained (2225, 21162) features. The vector area (2225, 21162) includes documents.

Table 4.3 Parameters used for NB 2225 document.

Parameters	Value	Details
Alpha	1	Using Laplace for smoothing parameter
Fit prior	True	Whether to learn class prior probabilities
Class prior	None	Prior probabilities of the classes
Binarize	0	The threshold for binarizing of sample features used by Multinomial
Norm	False	The second normalization of the weights is performed by complement
Var smoothing	1e – 9	Gaussian model in practice find, if the ratio of data variance between words is too small, it will cause numerical error. To address this problem, we artificially boost the variance

Table 4.4 Classification accuracy on the evaluation set using TF with 2225 document.

Category	NB Accuracy
business	70%
sport	75%
tech	50%
Politics	40%
entertainment	60%

Here, for training and testing purposes the data is split according to the percentage value that is used to divide the document. With regard to training, the hash value was 60%, that is, 2225 documents, and for the test, it was 40%, that is. Table 4.5 provides a comparison of the results of all models NB in TF.

```
x_train,x_test,y_train,y_test=train_test_split(TF,lable,test_size=40,random_state=60)
```

Figure 4.1 The code of training and testing in the first trial.

Table 4.5: Comparison of results in the first trial.

NB	Label	Precision	Recall	F1-score
Gaussian	business	1.00	1.00	1.00
	entertainment	1.00	0.91	0.95
	politics	0.83	1.00	0.91
	sport	1.00	1.00	1.00
	tech	1.00	1.00	1.00
Multinomial	business	0.90	0.82	0.86
	entertainment	1.00	0.82	0.90
	politics	0.67	0.80	0.73
	sport	1.00	1.00	1.00
	tech	0.50	1.00	0.67

In the above shown table 4.5 we have achieved 0.97 and 0.88 accuracy at the Naïve Bayes of Gaussian and Multinomial, respectively.

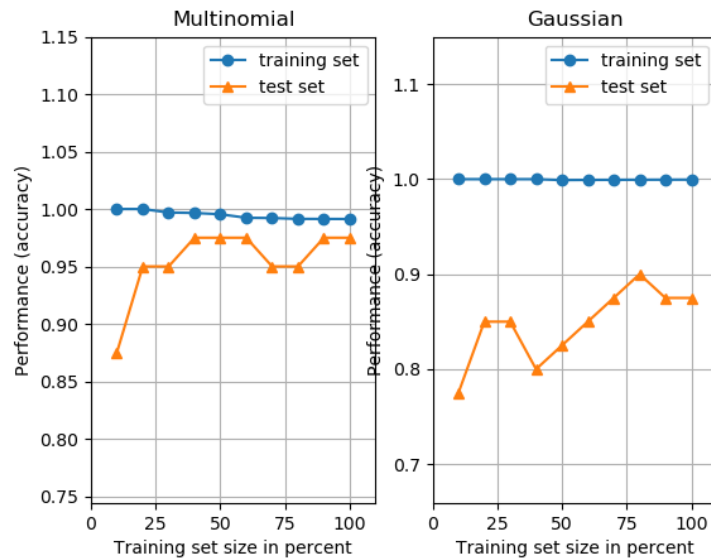


Figure 4.2: Training and testing of our model accuracy.

We also went through another training and testing trial with a new collection of documents splitting value. We have selected 70% (i.e., 2225 documents), 30% for training and testing respectively.

```
x_train,x_test,y_train,y_test=train_test_split(TF,table,test_size=0.30,random_state=70)
```

Figure 4.3: The code of training and testing in the second trial.

Table 4.6: Comparison of results in the second trial.

NB	Label	Precision	Recall	F1-score
Gaussian	business	0.99	0.99	0.99
	entertainment	0.99	0.90	0.94
	politics	0.96	0.99	0.98
	sport	0.99	1.00	1.00
	tech	0.93	0.98	0.95
Multinomial	business	0.88	0.91	0.89
	entertainment	0.89	0.91	0.90
	politics	0.90	0.84	0.87
	sport	0.98	0.97	0.98
	tech	0.91	0.93	0.92

In the above shown table 4.6 we have achieved 0.97 and 0.91 accuracy at the Naïve Bayes of Gaussian and Multinomial, respectively.

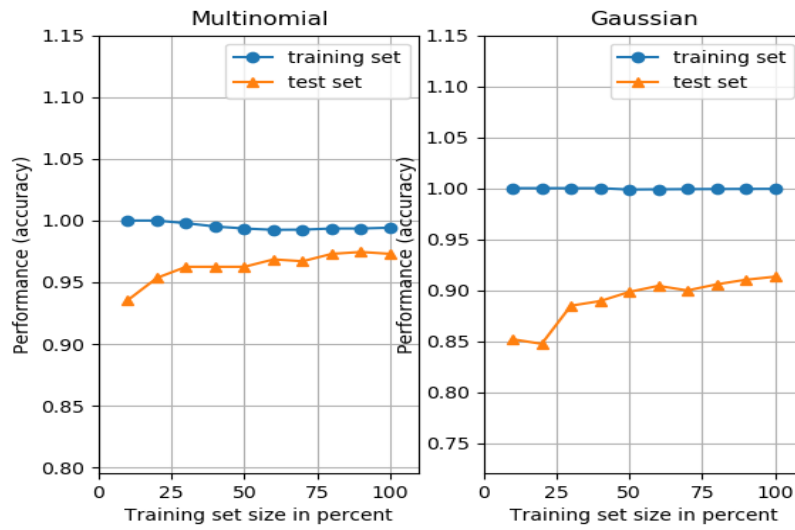


Figure 4.4: Training and testing of our model accuracy.

At the end I change the number of training and test 2 time each time I notice a different result of accuracy, in the first time the result was 0.97 and 0.88 and in the second time the accuracy was 0.97 and 0.91, the accuracy was politics, business, sport, technology, and entertainment.

The best accuracy I noticed was in the second time when the training was 30% and the test 70%.

In total (2225, 21162) documents downloaded from the Internet were used according to our experiments. Of these documents, a right classified by an expert is assigned. The verification method was common for dividing documents into a training and a test collection. Web Doc system execution was measured using well-known accuracy, recall, and F-metrics. for all tests, all filter indices were filtered conferring to their metering . for arbitrage, all metering were normalized in the domain from 0 to 1. Then, with a step of 0.1, a front from 0 to 0.9 was set. The testing effect for the two naive Bayesian samples is obtained in table 1. The arbitrage results for various favor processes are shown in table 2. Table 1 and Table 2 show the experimental results for two methods of selecting features. The classification of the replacement rely document on the collection of words created from the training of the document. therefore, the digit of coaching is documented such that seriose information about term collection can not define the category of the replacement document. A large digit of term collection from training documents decreases the likelihood of errors when classifying a document replacement. In the course of this study, together with access were executed as previously talk. The tests were completed using the over data sets in combination with the algorithm education of the naive Bayes classifier. Performance metrics are based on the precision, accuracy, and recall measured that were sitting in former parts.

4.7 EXPERIMENTS USING AN EVALUATION SET

Validation was used to determine the typical NB performance for text categorization and styling of training sets that provide the simplest capacity. This test confirming a detail and independently constructed estimate collection, which is designed to measure NB performance in a group of documents that have never been presented to the classifier. For this, we additionally neatly collected (2225, 21162) documents from BBC.net for each of the 5 predefined categories. For each category, we selected documents that best reflect the variability within the category.

We are requesting this collection of documents due to recruitment pricing. This set is presented to the classifier for categorization. For testing on the evaluation kit, trained NB classifiers are used. for each category, we use the NB classifier, which was trained using the training set, which provided the simplest classification accuracy of the category in verification experiments.

In our status, we employed the entire collection as a training (2225) appeared by 2225 texts, since the simplest verification precision was acquired during the experiment with a single pass with 2225 terms. Table 4.4 summarizes the NB performance results when tested using an estimated kit. The results showed higher performance for sports and, therefore, business class with a classification accuracy of more than 70%. Other categories vary from 40% to 60%. typical accuracy in all categories is 62%. The results gained in the experiment with the evaluation set vary depending on the effects gained in the verification experiments.

4.8 SUMMARY

Experimental results were analysed, interpreted, compared, validated, evaluated and discussed thorough . Besides, various analyses and discussions of the results are reported. The strengths and weaknesses of the proposed technology also are highlighted. Finally, the proposed method was created to be significantly better than recent works during this field.

5. CONCLUSION AND FUTURE WORK

5.1 INTRODUCTION

Its chapters minarets a compendium of the main returns and definitions and outlook function of this research.

Section 5.2 presents the conclusion of the study, and section 5.3 provides recommendations for future work.

5.2 RESEARCH CONCLUSION

The main objective of this research is to strengthen the organization of documents , because of the function of construct records into files. Managed algorithms suppose that the class supposes or scale of the term database is formerly famous. They order a training collection of records with checks and resumption a mission that maps records to prespecify division distances. awareness the temple of categories in the improvement and therefore creating rightly named training sets is extremely difficult. handbook sorting of documents, as you know, is a high-class and time-consuming task.

In this study, we focus on analysing BBC documents using two NB methods. (2225, 21162) documents collected from the BBC website included many areas (business, entertainment, politics, sports, and technology) for mood analysis. They controlled mood analysis using a paper pulp method using naive Bayes classifiers. For the analysis of moods, two stages were used: pre-processing, extraction of attributes, data separation, and, ultimately, classification. Here we used naive Bayesian methods to classify data. The naive Bayes complement categorize indicates a maximum precision of 97%. How the reception functions.

5.3 RECOMMENDATIONS FOR FUTURE WORK

The assumption for future study are as follows: in modern years, a floppy domain of ready knowledge employs has been provided that will be used with smartphones and portable mobile (tablet) devices, as this functionally categorize the feature of the order data. Records categorize is an important agent for classifying documents of various types created in the era of great data. A so communal type of record completion is the naive Bayesian categorization. The naive Bayesian planner has founded a showing classification that changes greatly depending on the parent tactics applied in the records. Over the past 20 years, several discuss have tested with the

function of classifying term documents employ machine learning. The major goal of this function is to increase the capacity and precision of the categorize. naive Bayes method we applied for works fully even with great datasets. we are planning to expand this study by contract the NB function chosen method experienced in this article with another function chose the method, such reference amplification, basic strain permission (PCA). In addition, we will research another issue related to the opinion of the field, such as refusal, the irony. In addition to using feature selection.



REFERENCES

- [1] J. P. Walsh and T. Bayma, “Computer networks and scientific work,” *Soc. Stud. Sci.*, vol. 26, no. 3, pp. 661–703, 1996.
- [2] M. R. Keyvanpour and M. B. Imani, “Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms,” *Intell. Data Anal.*, vol. 17, no. 3, pp. 367–385, 2013.
- [3] A. Sarlan, C. Nadam, and S. Basri, “Twitter sentiment analysis,” in *Proceedings of the 6th International Conference on Information Technology and Multimedia*, 2014, pp. 212–216.
- [4] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, “Sentiment analysis: A review and comparative analysis of web services,” *Inf. Sci. (Ny)*, vol. 311, pp. 18–38, 2015.
- [5] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis,” *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, 2013.
- [6] G. Paltoglou, “Sentiment Analysis in Social Media,” pp. 3–17, 2014, doi: 10.1007/978-3-7091-1340-0_1.
- [7] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, p. e1253, 2018.
- [8] A. Reyes, P. Rosso, and T. Veale, “A multidimensional approach for detecting irony in twitter,” *Lang. Resour. Eval.*, vol. 47, no. 1, pp. 239–268, 2013.
- [9] V. K. Bongirwar, “A survey on sentence level sentiment analysis,” *Int. J. Comput. Sci. Trends Technol.*, vol. 3, no. 3, 2015.
- [10] M. J. Maña-López, M. De Buenaga, and J. M. Gómez-Hidalgo, “Multidocument summarization: An added value to clustering in interactive retrieval,” *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 215–241, 2004.
- [11] Y. Wang, L. Kung, and T. A. Byrd, “Big data analytics: Understanding its capabilities

- and potential benefits for healthcare organizations,” *Technol. Forecast. Soc. Change*, vol. 126, pp. 3–13, 2018.
- [12] A. Karami, L. S. Bennett, and X. He, “Mining public opinion about economic issues: Twitter and the us presidential election,” *Int. J. Strateg. Decis. Sci.*, vol. 9, no. 1, pp. 18–28, 2018.
- [13] T. Kindt and H.-H. Müller, *The implied author: concept and controversy*, vol. 9. Walter de Gruyter, 2008.
- [14] N. Leavitt, “Mobile phones: the next frontier for hackers?,” *Computer (Long. Beach. Calif.)*, no. 4, pp. 20–23, 2005.
- [15] Y. Wu *et al.*, “Surfactant-Assisted Phase-Selective Synthesis of New Cobalt MOFs and Their Efficient Electrocatalytic Hydrogen Evolution Reaction,” *Angew. Chemie Int. Ed.*, vol. 56, no. 42, pp. 13001–13005, 2017.
- [16] R. Moraes, J. F. Valiati, and W. P. G. Neto, “Document-level sentiment classification: An empirical comparison between SVM and ANN,” *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, 2013.
- [17] Z.-S. Wu *et al.*, “Graphene anchored with Co₃O₄ nanoparticles as anode of lithium ion batteries with enhanced reversible capacity and cyclic performance,” *ACS Nano*, vol. 4, no. 6, pp. 3187–3194, 2010.
- [18] P. Waila, V. K. Singh, and M. K. Singh, “Evaluating machine learning and unsupervised semantic orientation approaches for sentiment analysis of textual reviews,” in *2012 IEEE International Conference on Computational Intelligence and Computing Research*, 2012, pp. 1–6.
- [19] R. Suchdev, P. Kotkar, R. Ravindran, and S. Swamy, “Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach,” *Int. J. Comput. Appl.*, vol. 103, no. 4, 2014.
- [20] L. Augustyniak *et al.*, “Simpler is better?: lexicon-based ensemble sentiment classification beats supervised methods,” in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2014, pp. 924–929.

- [21]M. Z. Asghar, S. Ahmad, M. Qasim, S. R. Zahra, and F. M. Kundi, "SentiHealth: creating health-related sentiment lexicon using hybrid approach," *Springerplus*, vol. 5, no. 1, p. 1139, 2016.
- [22]M. Bekkali and A. Lachkar, "Arabic Tweets Categorization Based on Rough Set Theory," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, 2014.
- [23]S. Fatima and B. Srinivasu, "Text Document categorization using support vector machine," *Int. Res. J. Eng. Technol.*, vol. 4, no. 2, pp. 141–147, 2017.
- [24]M. Thunnissen and P. Van Arensbergen, "A multi-dimensional approach to talent: An empirical analysis of the definition of talent in Dutch academia," *Pers. Rev.*, vol. 44, no. 2, pp. 182–199, 2015.
- [25]T. Burns, *The BBC: Public institution and private world*. Springer, 2016.
- [26]V. Belair-Gagnon, *Social media at BBC news: The re-making of crisis reporting*. Routledge, 2015.
- [27]A. Montoyo, P. MartíNez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments." Elsevier, 2012.
- [28]S. R. Veil, T. Buehner, and M. J. Palenchar, "A work-in-process literature review: Incorporating social media in risk and crisis communication," *J. contingencies Cris. Manag.*, vol. 19, no. 2, pp. 110–122, 2011.
- [29]S. Kausar, X. Huahu, M. Y. Shabir, and W. Ahmad, "A Sentiment Polarity Categorization Technique for Online Product Reviews," *IEEE Access*, 2019.
- [30]V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: a survey of techniques," *arXiv Prepr. arXiv1601.06971*, 2016.
- [31]S. Bhattasali, "Correlations Between Sentiment Analysis of Movie Tweets , Film Critics Reviews , and Box Office Earnings," 2013.
- [32]R. S. Karp and B. S. Schlessinger, *The basic business library: core resources*. Greenwood Publishing Group, 2002.
- [33]E. Haddi, "Sentiment analysis: text, pre-processing, reader views and cross domains."

Brunel University London, 2015.

- [34] S. Paquette, P. T. Jaeger, and S. C. Wilson, "Identifying the security risks associated with governmental use of cloud computing," *Gov. Inf. Q.*, vol. 27, no. 3, pp. 245–253, 2010.
- [35] N. Habash, "Arabic morphological representations for machine translation," in *Arabic computational morphology*, Springer, 2007, pp. 263–285.
- [36] I. Djurović, M. Urlaub, L. J. Stanković, and J. F. Böhme, "Estimation of multicomponent signals by using time-frequency representations with application to knock signal analysis," in *2004 12th European Signal Processing Conference*, 2004, pp. 1785–1788.
- [37] G. K. Smyth and T. Speed, "Normalization of cDNA microarray data," *Methods*, vol. 31, no. 4, pp. 265–273, 2003.
- [38] G. Forman, "Feature selection for text classification," *Comput. methods Featur. Sel.*, vol. 16, pp. 257–274, 2007.
- [39] L. Albraheem and H. S. Al-Khalifa, "Exploring the problems of sentiment analysis in informal Arabic," in *Proceedings of the 14th international conference on information integration and web-based applications & services*, 2012, pp. 415–418.
- [40] J. W. Reed, Y. Jiao, T. E. Potok, B. A. Klump, M. T. Elmore, and A. R. Hurson, "TF-ICF: A new term weighting scheme for clustering dynamic data streams," in *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*, 2006, pp. 258–263.
- [41] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and word2vec for text classification with semantic features," in *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 2015, pp. 136–140.
- [42] D. Chopra, "Sentiment Analysis of News Headlines using NaAve Bayes Classifier," *Counc. Res. Dev. Enterp.*, 2015.
- [43] J. L. Berral *et al.*, "Towards energy-aware scheduling in data centers using machine learning," in *Proceedings of the 1st International Conference on energy-Efficient Computing and Networking*, 2010, pp. 215–224.

- [44]H. Kim and S. Lee, “An intelligent information system for organizing online text documents,” *Knowl. Inf. Syst.*, vol. 6, no. 2, pp. 125–149, 2004.
- [45]S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, 2007.
- [46]P. Y. Pawar and S. H. Gawande, “A comparative study on different types of approaches to text categorization,” *Int. J. Mach. Learn. Comput.*, vol. 2, no. 4, p. 423, 2012.
- [47]K. P. Murphy, “Naive bayes classifiers,” *Univ. Br. Columbia*, vol. 18, p. 60, 2006.
- [48]M. Al-diabat, “Arabic text categorization using classification rule mining,” *Appl. Math. Sci.*, vol. 6, no. 81, pp. 4033–4046, 2012.
- [49]J. C. Griffis, J. B. Allendorfer, and J. P. Szaflarski, “Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans,” *J. Neurosci. Methods*, vol. 257, pp. 97–108, 2016.
- [50]J. Wu, Z. Cai, and X. Zhu, “Self-adaptive probability estimation for naive bayes classification,” in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.
- [51]V. Gourmet, “Real-time hand pose estimation on a smart-phone using Deep Learning.” 2019.
- [52]M. J. Pazzani and D. Billsus, “Content-based recommendation systems,” in *The adaptive web*, Springer, 2007, pp. 325–341.
- [53]G. M. Foody, “Local characterization of thematic classification accuracy through spatially constrained confusion matrices,” *Int. J. Remote Sens.*, vol. 26, no. 6, pp. 1217–1228, 2005.
- [54]S. M. Vohra and J. B. Teraiya, “A comparative study of sentiment analysis techniques,” *J. JIKRCE*, vol. 2, no. 2, pp. 313–317, 2013.
- [55]M. I. Khalaf, D. Al-jumeily, and A. L. Eds, *Applied Computing to Support Industry*, vol. 2. Springer International Publishing, 2019.
- [56]P. Kumbhar and M. Mali, “A survey on feature selection techniques and classification

- algorithms for efficient text classification,” *Int. J. Sci. Res.*, vol. 5, no. 5, p. 9, 2016.
- [57] T. E. Oliphant, “Python for scientific computing,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 10–20, 2007.
- [58] N. Ari and M. Ustazhanov, “Matplotlib in python,” in *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)*, 2014, pp. 1–6.
- [59] F. Perez, B. E. Granger, and J. D. Hunter, “Python: an ecosystem for scientific computing,” *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 13–21, 2010.

