## REVIEW CLASSIFICATION USING NLTK AND

**SPACY**

***AMini Project report Submitted to Jawaharlal Nehru Technological University, Kakinada, in the partial fulfillment for the award of the Degree in***

### MASTER OF COMPUTER APPLICATIONS

***Submitted by***

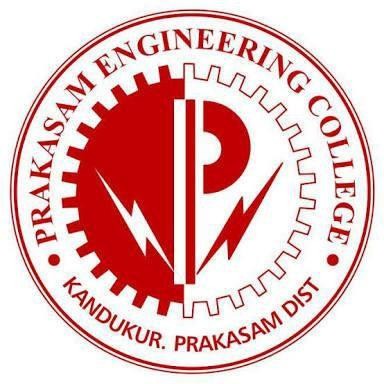
##### SIRIGIRI SAIMANIKANTA

##### (22F91F0059)

***Under the noble guidance of***

**Mr.MM RAYUDU,M.Tech(Ph.D)**

**Associate Professor**

****

## PRAKASAM ENGINEERING COLLEGE

***(An ISO 9001-2008 & NAAC Accredited Institution)* (Affiliated to Jawaharlal NehruTechnological University, Kakinada) O.V.ROAD, KANDUKUR-523105, A.P.**

**2023-2024**

# PRAKASAM ENGINEERINGCOLLEGE

***(An ISO 9001-2008 & NAAC Accredited Institution)***

**(Affiliated to Jawaharlal Nehru Technological University, Kakinada)**

**O.V.ROAD,KANDUKUR-523105,A.P.**



### DEPARTMENT OF

**MASTER OF COMPUTER APPLICATIONS BONAFIDECERTIFICATE**

*This is to certify that the mini project entitled* **“REVIEW CLASSIFICATION USINGNLTKANDSPACY”***isabonafideworkof***SIRIGIRISAIMANIKANTA(22F91F0059)** *in the partial fulfillment of the requirement for the award of the degree in***MASTER OF COMPUTER APPLICATIONS** *for the academic year 2023-2024. Thiswork is doneunder my supervision and guidance.*

### Signature of the GUIDE Signature of the HOD

##### Mr.M.M.RAYUDU Mr.M.M.RAYUDU

**M.Tech(Ph.D) M.Tech,(Ph.D)**

**Signatureof theExternal Examiner**

# DECLARATION

I do here by declare that the projectwork entitled**“REVIEW CLASSIFICATIONUSING NLTK AND SPACY”** is a genuine work carried out by me under the guidanceof **Mr.M.M.RAYUDU M.Tech (Ph.D)** in partial fulfillment for the award of the degree of**“Master of Computer Applications”** of **Jawaharlal Nehru Technological University, Kakinada.**

##### SIRIGIRI SAIMANIKANTA

**(22F91F0059)**

# ACKNOWLEDGEMENT

I feel to render my thankful acknowledgement to the following distinguished personalities, who stretched their helping hand to me, in completing my mini project work.

I am very grateful and my sincere thanks to our secretary & correspondent **Dr.K.RAMAIAH** of **PRAKASAM ENGINEERING COLLEGE** for giving this opportunity.

I hereby, express my regards and extend my gratitude to our PRINCIPAL, **Dr.CH.RAVI KUMAR** , for giving this opportunity to do the thesis as a part of our course.

I express my deep sense of gratitude to **Mr.M.M.RAYUDU, M.Tech(Ph.d),** Head of the Department, **Department of MCA** for having shown keen interest at every stage of development of our thesis and guiding us in every aspect.

And I am thankful to my Internal Guide **Mr.M.M.RAYUDU, M.Tech(Ph.d)** who has channeled my thoughts and timely suggestions.

I would also like to thank all my Faculties in Prakasam Engineering College for their constant encouragement and for being a great group of knowledgeable and cooperative people to work with.

##### SIRIGIRISAIMANIKANTA

**(22F91F0059)**

**TABLEOFCONTENTS**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CHAPTER**  **ABSTRACT** | | | | | **PAGENO.** | | |
| **1.INTRODUCTION** | | | | 1 | | | |
| **2. LITERATURE SURVEY**  2.1 General Overview of Sentiment Analysis and Classification  2.2 Factors that affect sentiment analysis and classification | | | | 4  4  8 | | | |
| **3. SYSTEM ANALYSIS** | | | | 8 | | | |
| 3.1 | | | Existing System | 9 | | | |
| 3.2 | | | Problem Statement | 9 | | | |
| 3.2 | | | Proposed System | 9 | | | |
| 3.3 | | | Feasibility Study | 9 | | | |
|  | | | 3.4.1 Economical Feasibility | 9 | | | |
|  | | | 3.4.2 Technical Feasibility | 10 | | | |
|  | | | 3.4.3 Social Feasibility | 10 | | | |
| 3.5 | | | Software Requirements Specifications | 11 | | | |
|  | | | 3.5.1 Software Requirements | 11 | | | |
|  | | | 3.5.2 Hardware Requirements | 11 | | | |
| **4. SYSTEM DESIGN** | | | | **12** | | | |
| 4.1 | | | System Architecture | 13 | | | |
|  | | | 4.1.1 Data Collection | 13 | | | |
|  | | | 4.1.2 Data Visualisation | 14 | | | |
|  | | | 4.1.3 Data Pre-Processing | 14 | | | |
|  | | | 4.1.4 Extracting Features | 14 | | | |
|  | | | 4.1.5 Data Splitting | 14 | | | |
|  | | | 4.1.6 Model Building | 15 | | | |
|  | | | 4.1.7 Predicting Target | 15 | | | |
| **5. SYSTEM IMPLEMENTATION** | | | | **16** | | | |
| 5.1 | | | Dataset | 17 | | | |
| 5.2 | | | Data Pre-Processing | 18 | | | |
|  | | 5.2.1. Cleaning The Data | | | 18 | |
|  | | 5.2.2. Stemming And Lemmatization | | | 19 | |
|  | | 5.2.3 Tokenization | | | 20 | |
|  | | 5.2.4 Normalization | | | 21 | |
|  | | 5.2.5 Stop Words | | | 21 | |
| 5.3 | | Data Visualisation | | | 23 | |
| 5.4 | | Extracting Features | | | 23 | |
|  | | 5.4.1 Tf-Idf | | | 25 | |
|  | | 5.4.2 Bag Of Words | | | 26 | |
| 5.5 | | Data Splitting | | | 27 | |
| 5.6 | | Model Building | | | 28 | |
|  | | 5.6.1 Decision Tree Classifier | | | 30 | |
|  | | 5.6.2 Random Forest Classifier | | | 31 | |
| **6. SYSTEM TESTING** | | | | | **41** | |
| 6.1 | | Model Evaluation | | | 42 | |
| **7. RESULT ANALYSIS** | | | | | **46** | |
| **8.CONCLUTION** | | | | | **48** | |
| **9.FUTURE ENHANCEMENT** | | | | | **50** | |
| **10. REFERENCES** | | | | | **52** | |

**ABSTRACT**

The rise of the internet has made online reviews an increasingly valuable andsignificant source of information for people. Consequently, there has been a surgeof interest in automatic review mining and summarizing as a prominent researcharea. "Text Classification" stands out as a crucial task in Natural LanguageProcessing (NLP). It involves categorizing text strings or documents into differentgroups based on their content. Some instances of text classification includedeterminingaudiencesentiment fromsocialmedia,identifyingspamand non-spamemails, tagging customer queries automatically, and classifying blog posts intovariouscategories.Thisparticularprojectconcentratesonaspecificdomain,whichis Movie Reviews. It employs a multi-knowledge-based approach that combinesstatistical analysis with movie knowledge. The dataset utilized in this project is theMovie Review Dataset, sourced from the IMDB movie reviews dataset. Thisdataset comprises thousands of positive and negative movie reviews. To ensureconsistency, the dataset undergoes various cleaning and preprocessing techniques,suchas convertingallreviewstoEnglishlanguage.The experimentalfindingsdemonstratetheeffectivenessofthemulti-knowledge-basedapproachinminingand summarizing movie reviews, thereby providing the audience with a conciseand comprehensivefinalassessmentofthemovie.

I

**CHAPTER1**

### INTRODUCTION

**1.1.INTRODUCTION**

ReviewClassification,alsoknownasSentimentAnalysis,involvescategorizingtheviewpointorreviewexpressedinatextusingtechniquesfrominformationretrieval andcomputational linguistics. Rather than focusing on the topic itself, the emphasis isplaced onthe significance of the review. Sentiment analysis utilizes methods such as natural languageprocessingandtext analyticsto extract subjectiveinformationfrom source materialslikereviews. Reviews play a crucial role in our decision-making process by providing valuableinsights. Online review sites and personal blogs leverage information technologies to gathersentiments about products or objects. The primary objective of Review Classification is todetermine the polarity of comments (positive, negative, or neutral) by identifying the featuresandcomponentsoftheobjectthathavebeencommentedonineachdocument.

IntherealmofReviewClassificationstudies,thefocusisontheeconomicimpactresulting from reviews, as well as concerns regarding privacy breaches. These studies examinethenatureofreviews,whichcanbecategorizedaseitherdirectreviewsorcomparativereviews. Direct reviews express sentiments towards specific targets such as products, events,topics, or individuals. For example, a direct review might state, "The climax of the movie isgripping."Ontheotherhand,comparativereviewshighlightthesimilaritiesordifferences

betweenmultipleobjectsandoftenestablishanorderofpreference.Forinstance,acomparative review might state, "Movie X is better than movie Y." Comparative reviews canfurther be classified into different types, including Non-equal Gradable (comparisons using"lessthan"),Equative(comparisonsusing"same"),andSuperlative(comparisonsusing"longest").As of now, there are eight teams that compete with one another in a double round-robin fashion during the league stage. After the league stage, the top four teams in the leaguepointstablequalifytotheplayoffs.

Review Classification can be conducted either at the document level or at the sentencelevel. At the sentence level, there are two tasks involved: subjective classification and objectiveclassification.

Objective classification is applied to sentences that provide factual information withoutanypersonalopinion. Forexample,"Iwatchedamoviefew daysago."

Subjective classification, on the other hand, is used for sentences that express personalopinions or feelings. These sentences can be further classified as positive or negative. Forinstance, "It is such a nice movie" would be classified as positive, while "The movie wasboring"would beclassified asnegative.

At the documentlevel, the classification isbasedon theoverall sentimentexpressed bythe review holder. The document, such as a review, is classified as either positive or negativebasedonthesentimentconveyed.

In summary, Review Classification can be performed at either the document level or thesentence level. At the sentence level, subjective and objective classifications are carried out,while at the document level, the overall sentiment of the review determines the classification aspositiveornegative.

Assumption:Eachdocumentfocusesonasingularobjectandcontainsopinionsfromasoleopinionholder.Forinstance,thumbs-upor thumbs-down,starratings (1star, 2stars,3stars...).

Reviewscan also be basedon specificfeatures,asexemplifiedin thisexample."Thisfilmisincredibly misunderstood, to the point where it's not even amusing. If you're considering watching itfor the action scenes, I advise against it. This film delves into the effects and trauma that survivorsmust endure. Even the detectives are searching for the same answer we all are... WHY? The twoleadingladiesdeliverfantasticperformances,showcasinghowthoseweoverlookareimpactedbythe very same things that affect us. Yes, the language can be harsh at times, but it suits the charactersperfectly. There are a few loose ends or unanswered questions, but that's common in all movies. Themain issues are addressed, and thisfilm makes a significant statement about how adults feel aftersuchmajorincidents.Ihighlyrecommenditforteenagers and adults..."

Each aspect of the product is categorized, and an overall sentiment is evaluated. This project presentsasurveyonvarious methodsofsentimentanalysisfoundinliteraturepertainingtoproductreviews.

### CHAPTER2

1. **LITERATURESURVEY**

The primary objective of this project is to determine the underlying sentiment of a movie reviewbased on its textual information. Our goal is to classify whether a person enjoyed or disliked themovie based on their review. This is particularly valuable when filmmakers want to assess the overallperformance of their movie by analyzing the reviews provided by critics and viewers. The results ofthis project can also be utilized to create a movie recommender system, which suggests movies toviewersbased on their previousreviews. Additionally, thisproject can helpidentify groupsofviewers with similar movie preferences. As part of this project, we will study various techniques forextracting features from text, such as keyword spotting, lexical affinity, and statistical methods, andanalyzetheirrelevancetoourproblem.Furthermore,wewillexploredifferentclassificationtechniques and evaluate their performance with different types of feature representations. Ultimately,we will draw a conclusion regarding the most accurate combination of feature representations andclassificationtechniquesfor thecurrentpredictivetask.

The dataset was initially analyzed by researchers at Stanford University, who employed unsupervisedlearning techniques to group words with similar meanings and generate word vectors. These wordvectors were then utilized in various classification models to determine the sentiment polarity of thereviews.Thismethodologyprovesparticularlyvaluablewhendealingwithdatathatcontainsextensivesentiment-relatedcontentandissusceptibletosubjectiveinterpretationsofwordassociations. Additionally, Bo Pang and Peter Turnkey have made significant contributions in thefield of polarity detection for both movie and product reviews. Their work also encompasses thedevelopment of a multi-class classification system for reviews and the prediction of reviewer ratingsformoviesand products.

These studies explored the utilization of various classification techniques such as Random Forestclassifier, Linear Regression, Decision tree, MLP classifier, and Multinomial naïve Bayes for thepurpose of categorizing reviews. Additionally, these works also focused on the implementation ofdifferent featureextractionmethods.Asignificant aspect highlightedintheseresearchpaperswastheexclusion of a neutral category during classification. This decision was based on the assumption thatneutral texts tend to be located near the boundary of binary classifiers and pose a disproportionatechallenge in terms of classification. Currently, there are numerous sentiment analysis tools andsoftware available either for free or under commercial license. The rise of microblogging has led tothe widespread use of sentiment analysis in order to analyze public sentiments and derive meaningfulinsights from them. One notable example of its application was the utilization of Twitter data tocomprehendthepoliticalsentimentof individualsinrelationtotheGermanFederalelections.

InSection2.1,wepresentanoverviewofhowsentimentanalysisandclassificationworkingeneral.InSection2.2,wediscusssomefactors thataffectsentimentanalysisandclassification

* 1. **2.1 GeneralOverviewofSentimentAnalysisandClassification**

The classification of sentiments can be done at different levels, namely the word, sentence, ordocument level. However, the focus of this particular project is on document-level sentimentclassification. Document-level sentiment classification involves determining whether a given set ofrelateddocuments,whichare opinionatedinnature,expressapositive ornegative opiniontowardsaparticular object. In the existing research conducted in this field, it is assumed that the opinionateddocument (such as a movie review) contains opinions pertaining to a single object. This assumptionis generally valid for customer reviews of products and services. However, it may not hold true forforum and blog posts, as these types of posts can contain opinions about multiple products andservices.

Todeterminethesentimentof adocument,onepossibleapproachistoanalyzethesentimentofeachsentence individually and then aggregate the results to obtain the overall sentiment of the document.However, this method requires first classifying the sentiment of each word within the sentencesbefore performing the sentence-level analysis. A major challenge in document-level sentimentanalysisisthatnotallparts ofthedocumentcontributeequallytoinferringtheoverallsentiment.

Identifying the relevant sentences automatically poses its own learning difficulty. Some researchershave proposed automated methods for extracting meaningful sentences, while others have exploredalgorithmsthatconsiderallsentencesinthedocumenttoachieveimprovedexperimentaloutcomes.

**2.2 Factorsthataffectsentimentanalysisandclassification**

classifying a document as positive or negative based on the overall sentiment expressed by the author is considered a more challenging task compared to strict text classification. This is because opinions can often be conveyed in a more intricate manner, making it difficult to identify them solely based on individual words within a sentence or document. Merely analyzing the words used in a review may not provide an accurate classification of the sentiments expressed. Take, for instance, this review: "This film should be brilliant. It sounds like a great plot, the actors are first

grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance, however, it can't hold up".

The utilization of termslike "brilliant," "great," and "good" implies afavorable sentiment, leadingone to believe that determining the sentimentof areviewthrough a predefined setof keywordswould be straightforward. Nevertheless, Pang andLee's experimental findings revealed that theaccuracy of classification achieved through human-generated keyword lists was inferior to that ofkeywords generatedthroughmachinelearning techniques.

The review mentioned above contains an anaphor, such as the phrase "it can't hold up." The referentof"it"inthisphraseisunclear,whetheritpertainstothemovieorStallone'sperformance.Consequently, it becomes challenging to ascertain the overall sentiment of the review. Numerousfree-form reviews employ multiple anaphora, abbreviations, and exhibit a lack of capitalization,improperspelling, punctuation,and grammar.

The analysis and classification of opinions are influenced by several key factors. These factorsinclude the domain of the datasets, the size of the datasets being analyzed, the format of the datasets(whether they are labeled or unlabeled), and the quality of the dataset. The accuracy of sentimentclassification can be affected by the specific domain to which it is applied. This is because the samephrase can have different sentimentsin different domains. For example, the phrase"go read thebook" is likely to indicate positive sentiment in book reviews, but negative sentiment in moviereviews. Similarly, a review like "it is so easy to predict the next action..." may be negative sentimentforamovieplot,butpositivesentimentforapoliticalreview.Additionally,thevariationinvocabularies across different domains poses a challenge when applying classifiers trained on labeleddatafromonedomaintotestdatafromanother domain.

Depending on the size of datasets to be utilized, either manual or automatic, or both methodologies,could be employed. Nevertheless, it is always advisable to employ a combination of both approaches.Numerous experiments have revealed that even the poorest outcomes obtained from utilizing bothapproaches surpass the best outcomes achieved through manual approaches and certain automaticapproaches.Thepresenceoflabeleddataalsoenhancestheefficiencyofopinionclassification.Inits

absence, many researchers have resorted to the linguistic/semantic approach of constructing lexicons,which is exceedingly time-consuming and yet does not yield superior performance. Furthermore,lexicons are language and domain-specific, thereby further complicating the sentiment analysis andclassificationtask.

The sentiment classification performance is directly influenced by the quality of the dataset. As thereis currently no mechanism in place to ensure the quality of reviews, anyone can post anything on theinternet. Consequently, a significant numberof low-quality reviews and review spam exist. Bing etal. conducted a study on opinion spam and found that online reviews often consist of spam messages,including false or fabricated reviews, irrelevant reviews, and reviews that are not genuine but ratherstatements or questions. To mitigate the presence of spam in these reviews, it is crucial to implementeffective pre-processing techniques. However, it is worth noting that this particular research area hasnotreceivedmuchattentionthusfar.

### CHAPTER3

**3.SYSTEM ANALYSIS**

##### 3.1 EXISTINGSYSTEM

Thecurrentprocessofcategorizingreviews iscarriedoutmanuallybyreadingeachreview and assigning it a positive or negative label. However, this system is unable to handlelarge volumes of data within the given timeframe. Moreover, this approach is limited toprocessing structured data and the accuracy of existing technologies is incomparable to that ofmoderntechnologies.

##### 3.2 PROBLEMSTATEMENT

The current system experiences a longer duration for review classification due tomanualprocessing. Consequently,itbecomesexceedinglychallengingtogenerateresultsforthevastamountofdatathatexistsintoday's world.

##### 3.3 PROPOSEDSYSTEM

TheclassificationofMovieReviews intospecificcategoriescanbeaddressedandsimplifiedbyemploying supervised machine learning algorithms for text classification. This project aims todevelop a machine learning model that automates the process of categorizing Movie Reviewsinto either positive or negative sentiments. Various models such as Random Forest, DecisionTree, MLP,LogisticRegression,etc.willbetrainedusingIMDBdataobtainedfromtheofficialIMDB website. The model with the highest accuracy will be deemed the most suitable for thisproject.

### 3.4 FEASIBILITYSTUDY

In this phase, the project's feasibility is examined and a business proposal is presented,outliningabroadplanfortheprojectalongwithcostestimates.Theproposedsystemundergoesa thorough feasibility study during the system analysis stage. This is done to ensure that theimplementation of the proposed system does not impose any unnecessary burden on thecompany. It is crucial to have a clear understanding of the key requirements for the system inordertoconductacomprehensivefeasibilityanalysis.

Threekeyconsiderations involvedinthefeasibilityanalysisare

##### ECONOMICALFEASIBILITY

1. **TECHNICALFEASIBILITY**

##### SOCIALFEASIBILITY

* + 1. **3.4.1 ECONOMICALFEASIBILITY**

Thisinvestigationaimsto assessthefinancialconsequencesthat thesystemwillhaveonthe organization. The organization's capacity to allocate funds for the research and developmentof the system is restricted. Therefore, the expenses must be rationalized. Consequently, thesystem was developed within the allocated budget, primarily utilizing freely availabletechnologies.Theonlyexceptionwastheacquisitionofcustomizedproducts.

### 3.4.2 TECHNICALFEASIBILITY

This research is conducted to assess the technical viability, specifically the technical prerequisites of the system. It is crucial that any system created does not impose excessive strain on the existing technical resources. Otherwise, it will result in an increased burden on the client. Therefore, the developed system should have a moderate requirement, necessitating minimal or no alterations for its implementation.

### 3.4.3 SOCIALFEASIBILITY

The purposeofstudyingistoassessthe degree ofacceptance ofthe systembytheuser. This encompasses the user training process to ensure efficient utilization of thesystem.

##### 3.5 SoftwareRequirementSpecification

* + 1. **3.5.1 SoftwareRequirements**

Operating System : Windows, Linux or UNIXSoftware :Anaconda

IDE : Jupyter notebookEditor : Text EditorToolkit:NLTK

CodeLanguage : Python3.6

##### 3.5.2 HardwareRequirements

Processor:Inteli3,i5,i7andaboveHardDisk:100GB

RAM:4GBorabove

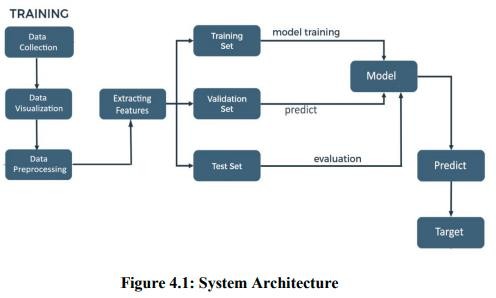
## CHAPTER4

**SYSTEMDESIGN**

## 4.1SYSTEMARCHITECTURE

In this section, we present a concise outline of the sentiment analysis and classification process.For a more comprehensive explanation, please refer to Chapter 5. Figure 1.1 illustrates the keystages involved in text classification using sentiment. The initial step, known as pre-processing,entails extracting the reviews or documents from a source dataset. Subsequently, varioustechniques are employed to clean the terms within each review. These terms, along with theircorrespondingsentimentscores,arethenstoredasfeaturevectors, whichserveas inputfor atextclassifier.

The subsequent phase entails employing a text classifier to categorize the chosen attributes aseither positive or negative. The feature vectors that have been stored are utilized as input for thetext classifier. If deemed necessary, feature selection can be implemented to diminish thequantityofattributes.Toachieveoptimaloutcomes, asequenceofiterativemeasures, referredtoas cross-validation, can be employed to assess the predictive model's practical performanceaccurately.



### 4.1.1 DataCollection

The AI department of Stanford University utilized the Large Movie Review Dataset to gather the dataset for this task. This dataset, consisting of 50,000 training examples, was collected from IMDb. Each review in the dataset is labeled with the movie's rating on a scale of 1-10.

##### 4.1.2 DataVisualisation

Data visualization refers to the visual depiction of data, where images are created to effectively convey the relationships within the data to the viewers. This is accomplished by employing a structured mapping between graphic elements and data values during the creation of the visualization. This mapping determines how data values will be visually represented, dictating the extent to which properties of the graphic elements, such as size or color, will alter to reflect changes in the data values.

##### 4.1.3 DataPre-processing

Data pre-processing plays a crucial role in the data mining process. The concept of "garbage in, garbage out" holds true in the realm of data mining and machine learning endeavors. Often, data collection methods lack strict control, leading to the inclusion of out-of-range values (e.g., Income: −100), implausible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, and more. Analyzing data without carefully addressing these issues can result in misleading outcomes. Therefore, ensuring the accuracy and quality of data is paramount before conducting any analysis. In many cases, data pre- processing serves as the most vital phase of a machine learning project, particularly in the field of computational biology.

**4.1.4 ExtractingFeatures**

Text processing involves dealing with words in a text as discrete and categorical features. To make this data usable for algorithms, we need to encode it in a specific manner. This process of transforming textual data into real-valued vectors is known as feature extraction. Two common techniques for numerical representation of text are Bag of Words and TF-IDF.

# 4.1.5 DataSplitting

**TrainingDataset**:Thesampleofdataused to fitthemodel.

**TestDataset**:Thesampleofdataused toprovideanunbiasedevaluationofafinal modelfitonthetrainingdataset.

**Validation Dataset**:The data sample is utilized to impartiallyassess the fitting of a model on the training dataset during theprocessoftuningthemodel'shyperparameters.However,asthemodel configuration incorporates the skill on the validationdataset,theevaluationbecomesincreasinglybiased.

4.1.6 ModelBuilding

The overall task in this project is for classification of reviews as favourable orunfavourable. Therefore, for this classification task we explored multiple classification models on above feature representations We employed a variety of models, starting from the basic Logistic Regression to the advanced SVM Classifier. Additionally, we utilized other classification models such as MLP Classifier, Decision Tree, and Random Forest Classifier. In addition to these, we trained the aforementioned feature representations on Naïve Bayes' Classifier, which is commonly used in text mining alongside Bag of Words and TF-IDF Modelling. Furthermore, we developed a model based on k-Nearest Neighbours to assess the similarity between reviews and classify them accordingly.

4.1.7 PredictingTarget

The feature of a dataset that you aim to comprehend better is referred to as the target variable. In supervised machine learning, historical data is utilized to identify patterns and establish connections between the target variable and other features in the dataset. The specific target variable will differ based on the business objective and the data that is accessible.

## CHAPTER5

##### 5. SYSTEMIMPLEMENTATION

**5.1 Dataset**

The AI department of Stanford University utilized the Large Movie Review Dataset for this task. This dataset consists of 50,000 training examples gathered from IMDb, where each review is labeled with a movie rating ranging from 1 to 10. Since sentiments are often binary, such as good/bad or happy/sad, we categorized these ratings as either 1 (like) or 0 (dislike) based on the ratings. If the rating was above 5, we inferred that the individual liked the movie; otherwise, they did not. Initially, the dataset was split into two subsets, each containing 25,000 examples for training and testing. However, we found this division to be sub-optimal due to the limited number of training examples, resulting in under-fitting. To address this, we redistributed the examples, allocating 40,000 for training and 10,000 for testing. Although this yielded better models, it also led to over-fitting on the training examples and poorer performance on the test set. Ultimately, we decided to employ Cross Validation, dividing the complete dataset into multiple folds with different samples for training and validation in each iteration. The final performance statistic of the classifier is then averaged over all results, resulting in improved accuracy across the board. A typical review text within the dataset appears as follows:.

I am a fan of television movies in general, and this particular one was quite impressive. The performances by the cast were consistently strong, and there were unexpected twists before each commercial break. It reminded me of a combination of "Medium" and "CSI."

Did anyone else notice that, under certain lighting, the daughter resembled a young Nicole Kidman? Is there any relation between them? I would definitely watch it again or rent it if it becomes available on video.

Dedee was fantastic in her role. I haven't seen her in many projects, but she portrayed her character convincingly. If you enjoy TV mystery movies, I recommend checking this one out if you get the chance.

In the text provided, we had to remove HTML tags like "<br>" as a pre-processingstepbeforeextractingfeatures.Weusedsimpleregularexpressionstoaccomplishthis.

Additionally, we made the text case-insensitive to facilitate counting word occurrences across all reviews and removed punctuation marks such as '!', '?' and others, as they do not contribute significant information and can have different connotations. These tasks were performed using standard Python libraries for text and string manipulation. Furthermore, we eliminated stop words from the text for certain feature extraction tasks, which will be discussed in more detail later on. It is important to note that we did not employ word stemming, as it can result in the loss of information by reducing words to their root forms.

##### 5.2 DataPre-processing

Prior to executing any algorithm, it is imperative to ensure the data is thoroughly cleansed, facilitating smoother processing. Furthermore, by proactively identifying and eliminating extraneous words, we can significantly enhance the precision of our algorithms.

**5.2.1 Cleaningthedata**

The presence of HTML tags in the IMDb reviews does not contribute to sentiment detection. Therefore, we have made the decision to eliminate all punctuation, including emoticons (which are already scarce). This simplifies our data processing. Additionally, we convert all text to lowercase. Furthermore, we utilize the Porter stemming algorithm to replace each word with its root form. As a result, words such as "cats" and "cat," or "running" and "run," are treated as identical. This approach has been proven to enhance the accuracy of sentiment analysis in classification tasks.

##### 5.2.2 Stemmingandlemmatization

Different forms of a word, such as organize, organizes, and organizing, are utilized in documents for grammatical purposes. Moreover, there exist families of derivationally related words with comparable meanings, like democracy, democratic, and democratization. In numerous instances, it appears advantageous for a search to yield documents containing any word from the aforementioned set.

Both stemming and lemmatization aim to reduce the various forms of a word, including inflectional and derivationally related forms, to a shared base form. For example:

am,are,is=>be

car,cars,car's,cars=>car

Theresultofthis mappingoftextwillbesomethinglike:

Theboy'scarsaredifferentcolors=>theboycarbediffercolor

Porter's algorithm is widely recognized as the most commonly used and highly effective algorithm for stemming English words. Although the complete algorithm is complex and extensive, we will provide a brief overview of its general approach. The algorithm comprises five distinct phases of word reductions, which are applied sequentially. Each phase incorporates specific conventions for selecting rules, such as choosing the rule from each rule group that is applicable to the longest suffix. In the initial phase, this convention is employed along with the following rule group:

##### 5.2.3 Tokenization

Tokenization is the process of dividing a character sequence into smaller units, known astokens,whilealsodiscardingspecificcharacterslikepunctuationmarks.Let'sconsideranillustrationoftokenization:

Input:Friends,Romans,Countrymen,lendmeyourears;.

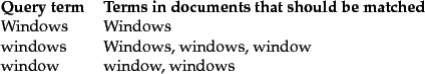
Output:



## 5.2.4 Normalization

Token normalization involves the normalization of tokens to ensure that matches occur even when there are slight differences in the character sequences of the tokens. The common approach to normalization is to create equivalence classes, typically named after one member of the set. For example, if the tokens "anti- discriminatory" and "antidiscriminatory" are both mapped to the term "antidiscriminatory" in both the document text and queries, searching for either term will retrieve documents that contain it.

Using mapping rules that eliminate characters such as hyphens offers the advantage of implicit equivalence classing. This means that the calculation of equivalence classes is not required beforehand, as the terms that become identical through these rules automatically form the equivalence classes. It is relatively simple to create rules that remove characters. However, since the equivalence classes are implicit, it is not always clear when it is necessary to add characters. For example, it would be challenging to determine that "antidiscriminatory" should be transformed into "anti-discriminatory".



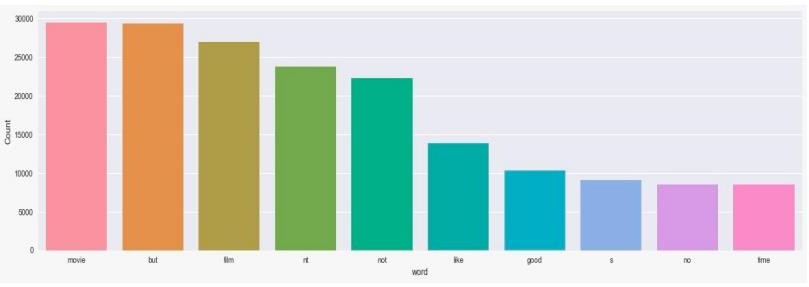
Instead of creating equivalence classes, an alternative approach is to establish relationshipsbetween unnormalized tokens. This technique can also be applied to curated lists of synonymslike car and automobile. These relationships between terms can be achieved in two differentways. The common approach involves indexing unnormalized tokens and maintaining a queryexpansion list that includes multiple vocabulary entries to consider for a specific query term. Asa result, a query term becomes a disjunction of several postings lists. The alternative approachinvolves expanding the index during its construction. For instance, when a document containstheterm"automobile," itisindexedunder"car"aswell(andviceversainmostcases).However,both of these methods are less efficient compared to equivalence classing since they requirestoringandmergingmorepostings.Thefirstmethodaddsaqueryexpansiondictionaryandrequires additional processing during query time, while the second method necessitates morestorage space for the postings. Traditionally, expanding the space required for postings lists wasconsideredadisadvantage.However,withthedecreasingcostsofmodernstorage,theincreasedflexibilityprovidedbydistinctpostingslistshasbecomemore appealing.

The flexibility of these methods surpasses that of equivalence classes due to the possibility ofoverlapping expansion lists without being identical. Consequently, an asymmetry in expansioncan occur. Figure 4 illustrates an example of how this asymmetry can be utilized. For instance, ifthe user inputs "windows," we aim to enable matches with the capitalized "Windows" operatingsystem.However,iftheuserenters"window,"itisnotfeasibletoconsideritamatch,despitethepossibilityofitalsomatchinglowercase "windows."

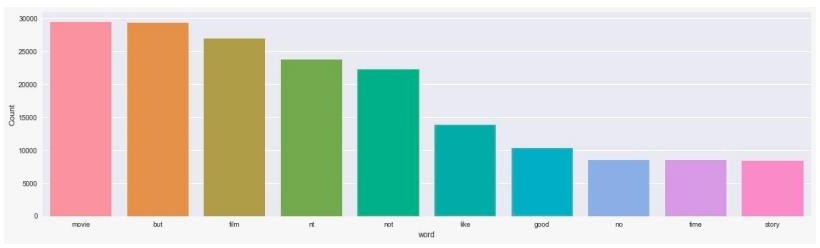
##### 5.2.5 StopWords



Occasionally, certain commonly used words that may seem insignificant in assisting withdocument selection for a user's needs are completely excluded from the vocabulary. These wordsare referred to as stop words. The typical approach for determining a stop list is to arrange theterms based on their frequency in the document collection and then select the most frequentterms, often filtered manually for their semantic relevance to the domain of the indexeddocuments. These selected terms form the stop list, and they are disregarded during the indexingprocess. An example of a stop list is displayed in the above Figure. By utilizing a stop list, thenumber of postings that a system needs to store is significantly reduced. Omitting the indexing ofstop words often has minimal impact on keyword searches that include terms like "the" and "by,"as they do not appear to be very useful. However, this is not the case for phrase searches. Forinstance, the phrase query "President of the United States," which contains two stop words, ismoreprecisethanusing"President"AND"UnitedStates."Iftheword"to"isincludedinthestoplist, the meaning of searches for flights to London may be lost. Similarly, searching for VannevarBush's article "As we may think" becomes challenging if the first three words are excluded, andthe system only searches for documents containing the word "think." Certain types of queries aredisproportionately affected by stop words. Some song titles and well-known verses consistentirely of words that commonly appear on stop lists (e.g., "To be or not to be," "Let It Be," "Idon't want to be..."). The trend in information retrieval (IR) systems has shifted from usingrelatively large stop lists (200-300 terms) to very small stop lists (7-12 terms) and, in some cases,eliminating stop lists altogether. Web search engines typically do not employ stop lists. Thedesign of modern IR systems has focused on leveraging language statistics to better handlecommonwords.



**Figure5.3:Beforeremovingstopwords**



**Figure5.4Afterremovingstopwords**

#### 5.3 DataVisualisation

Data visualization refers to the visual depiction of data, where images are created to effectively convey the relationships within the data to the viewers. This is accomplished by employing a structured mapping between graphic elements and data values during the creation of the visualization. This mapping determines how data values will be visually represented, dictating the extent to which properties of the graphic elements, such as size or color, will alter to reflect changes in the data values.

#### TagClouds

Tagcloudsarethemostbasicand widelyusedmethodofvisualizingtext. They displaytags in a spatial arrangement, with variations in size, color, and position reflecting thefrequency,categorization,orimportanceofeachtag.



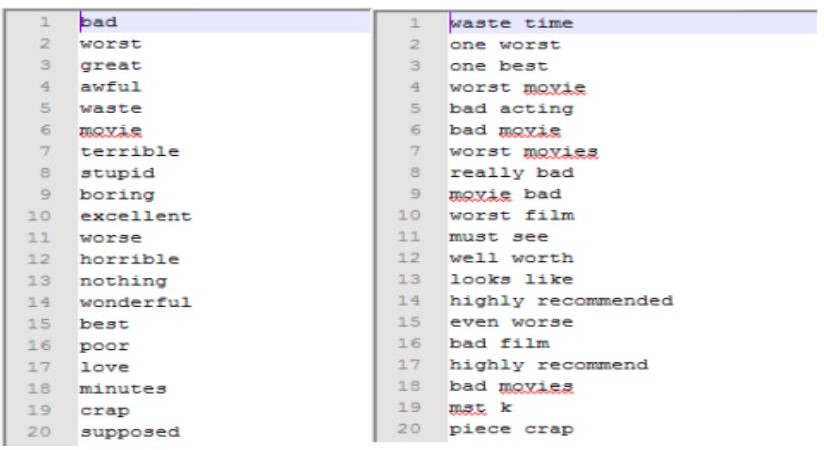
#### Figure 5.5: WordCloud

**5.4 ExtractingFeatures**

Text processing involves dealing with words in a text as discrete and categorical features. To make this data usable for algorithms, we need to encode it in a specific manner. This process of transforming textual data into real-valued vectors is known as feature extraction. Two common techniques for numerical representation of text are TF-IDF and Bag of Words.

**5.4.1 TF-IDF**

Tf-Idf, short for term frequency, inverse document frequency, is a valuable technique primarily employed in information retrieval to assess the significance of a keyword in a specific document within a corpus1. Reflecting on the term "movie," it led us to realize that there was still irrelevant content, prompting us to consider the possibility that a word could hold great importance in the entire collection of reviews but not in any individual review. Consequently, we required a method to determine the importance of words in relation to each review.



#### Figure5.6:Tf-Idf

**5.4.2 BagofWords**

The bag-of-words model is a method used in natural language processing to represent text. Instead of considering grammar or linguistic structures, this model treats a text as a collection of its individual words. After preprocessing the data, let's assume there are N unique words in the entire document. Additionally, suppose the document contains R reviews. Each review is represented by an N-dimensional feature vector. The entries in these vectors correspond to the N words in the document, and the feature values indicate the frequency of each word in the review. Alternatively, we can create a binary feature vector where 1 represents the presence of a word in the review, and 0 represents its absence. However, to handle the large size of the text document, we need to perform "feature selection" to reduce the dimensionality of the data points. After preprocessing and cleaning the data, we have a total of N = 78,767 words in the document. These words are ranked based on their importance, determined using the mutual information criteria discussed earlier. We select the top d = 5000 words to construct the feature vectors. It's worth noting that a larger number of words would likely yield more accurate results, but we are constrained by memory limitations and algorithm runtime.

#### 5.5 DataSplitting

Once the reviews have been cleaned and pre-processed, the data can be divided into three sets: train, validation, and test. These sets will be split in the ratio of 70%, 15%, and 15% respectively. The train set will be utilized for training the model, while the validation and test sets will be employed for making predictions and evaluating the model.

**TrainingDataset:**The sampleof dataused tofitthemodel.

**TestDataset:** Thesampleofdatausedtoprovideanunbiasedevaluationofafinalmodel fitonthetrainingdataset.

**Validation Dataset:**The data sample is utilized to impartially assess the fitting of amodel on the training dataset during the process of tuning the model's hyperparameters.However, asthemodelconfigurationincorporatestheskillonthevalidationdataset, theevaluation tendstobecomemorebiased.

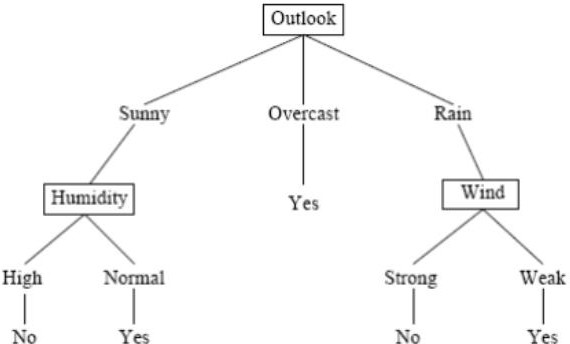
#### 5.6 ModelBuilding

The primaryobjectiveof this projectisto classifyreviewsas eitherfavorable orunfavorable. To accomplish this classification task, we explored various classificationmodels using the aforementioned feature representations. Our models ranged from thesimple Logistic Regression to the state-of-the-art SVM Classifier. Additionally, weutilizedotherclassificationmodelssuchasMLPClassifier,DecisionTree,andRandomForestClassifier.Inaddition tothesemodels,wealsotrainedthe above featurerepresentations on Naïve Bayes' Classifier, which is commonly used in text mining inconjunction with Bag of Words and TF-IDF Modelling. Furthermore, we developed amodel based on k-Nearest Neighbors to determine the similarity between reviews andclassify themaccordingly.

#### 5.6.1 DecisionTreeClassifier

A decision tree is a hierarchical structure where the internal nodes represent tests based on attributes, and the leaf nodes represent categories or classes. Each internal node tests a specific attribute, and each branch from a node corresponds to a value for that attribute.

The decision-making attribute is not predetermined, allowing for the selection of the attribute that provides the most information. Decision trees are not restricted to Boolean functions and can be applied to functions with categorical values as well.



#### Figure 5.9: DecisionTree

In the given example, the provided instances can be categorized by the values of the"outlook"attribute.The instancesare dividedbasedonattributes,and the decision foreach node is determined by selecting the attribute that provides the most information.Therefore,inthisexample,choosing"Outlook"astherootnodeyieldsthehighestamountof information atthat level.The edgesrepresent thepossible values oftheattributes, and the instances are divided accordingly among the child nodes. The tree canbea multi-waytree dependingonthepotentialattribute values.The selectionof attributesis based on aheuristicapproach,wherethe chosenattributeis expectedto provide thebestdivisionat a specificlevel.This approach has proventobesuccessfulin thepast.

#### 5.6.2 RandomForestClassifier

A Random Forest is a type of classifier that is composed of multiple tree-structured classifiers. Each tree in the forest independently generates random vectors and casts a vote for the most popular class at a given input x. The random vectors are generated independently of each other, with the same distribution, and a tree is created using the training set. In the case of random forests, a upper bound is derived to estimate the generalization error based on two parameters: the accuracy of the individual classifiers and the dependency between them.

Theerrorgeneralization forrandomforestcan bedividedintotwocomponents.Thesecomponents aredescribedasfollows:

1. Theeffectiveness ofeach individualclassifier withintheforest.
2. Thelevelofcorrelationbetweentheclassifiersbasedontherawmarginfunction.

To enhance the precision of the random forest, it is essential to reduce the correlationwhile preserving its robustness. According to Brieman's (2001) research, forests areconstructedbyrandomlyselectinginputsorcombinationsofinputsateachnodetogrowthe tree. This class of procedures possesses several advantageous qualities, which areoutlined asfollows:

1. Demonstratesgoodaccuracy,andinsomecases,evensuperiorperformance.
2. Exhibitsrelativeresilience tooutliersand noise.
3. Operatesatafaster pace compared toBaggingand Boostingtechniques.
4. Possessessimplicityandcanbe easilyparallelized.

Briemanhasproposedarandomizationapproachthatismoreeffectivewhenusedwithbagging or the random space method. The process of generating each tree in a randomforestinvolvesthefollowingsteps:

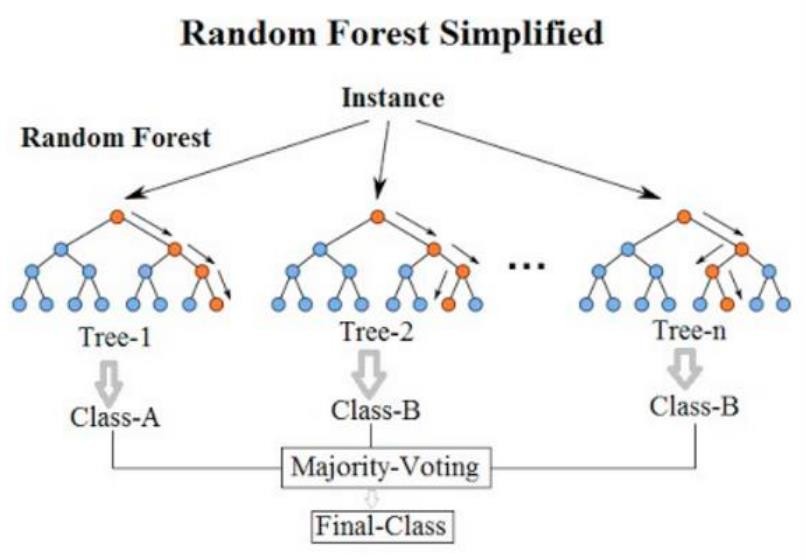
1. The trainingdataset contains Nrecords.
2. Nrecordsarerandomlysampledwithreplacement.
3. Thissampleddatasetisreferredtoas thebootstrapsample.
4. Ifthetrainingsetconsistsof Minputvariables,arandomselectionof m<<Minputsismade,andthe best spliton thesemattributes is usedtodivide thenode.
5. Thevalueof mremainsconstant throughoutthegrowth oftheforest.
6. Thetree isgrown toitsmaximumpossiblelevel.

Thereare a couple of justificationsfor employingtheBaggingapproach,asoutlinedbyBrieman (1994). Firstly, it appears that incorporating bagging in conjunction withrandomfeatures yieldsmore precise outcomes.Secondly,baggingcan be utilizedtocontinuouslyestimate generalizationerror,as well as thestrength and correlation.

When a new instance needs to be classified, the forest undergoes a process where eachtreeintheforestprovidesavotefortheclassificationoftheinstance.Thesevotesarethen combined and counted, and the classification with the highest number of votes isdeclaredas theclassification for thenew instance.Thisprocess isreferred toas theForestRIprocess.

To build the forest, a bootstrap sample is created by randomly selecting data withreplacement for each tree. As a result, one-third of the instances are left out and referredto as Out of Bag (OOB) data. Each tree in the forest has its own OOB data, which is usedto estimatethe errorofindividual trees,knownasOOB errorestimation.

Inaddition,randomforests alsoinclude built-infeatures for calculatingvariableimportanceandproximities.



**Figure5.10:RandomForest**

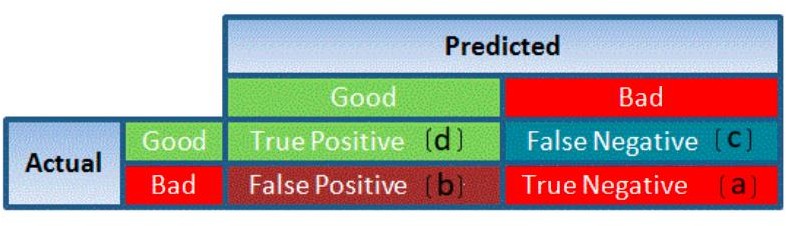
## CHAPTER6

#### 6.SYSTEMTESTING

**6.1 ModelEvaluation**

The evaluation of a model's performance is an essential step in the model buildingprocess. In order to determine the accuracy of predictions, it is necessary to comparethem with the actual values. This can be done by plotting the results and calculating thedistance betweenthepredictionsand actualvalues.Theaccuracyof thepredictionsincreases as the distance between them decreases. As this is a classification problem,there are several evaluationmetrics thatcan be used toassess theperformanceof ourmodels.

* **Precision**:We can gaina better understandingof accuracy byreferringtotheconfusionmatrix, which presents a tabular format comparing the actual and predicted values. Theconfusion matrix provides a visual representation of the data, allowing us to assess theaccuracy ofourpredictions.



#### Figure6.1:Accuracy

1. TruePositivereferstotargetsthataregenuinelytrue(Y)andhavebeenaccuratelypredictedastrue(Y).

TrueNegativereferstotargetsthatare genuinelyfalse(N)andhavebeenaccuratelypredictedasfalse(N).

FalsePositivereferstotargetsthatare genuinelyfalse(N)buthavebeenincorrectlypredictedastrue(T).

FalseNegativereferstotargetsthataregenuinelytrue(T)buthavebeen incorrectlypredictedasfalse(N).

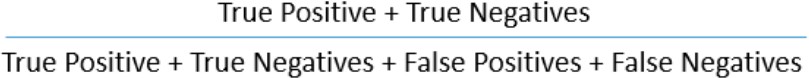
1. TruePositivedenotestargetsthatareindeedtrue(Y)andhavebeencorrectlypredictedassuch(Y).

TrueNegativedenotestargetsthatareindeedfalse(N)and havebeencorrectlypredictedasfalse(N).

FalsePositivedenotestargetsthatareindeedfalse(N)buthavebeenerroneouslypredictedastrue(T).

FalseNegativedenotestargetsthatareindeedtrue(T)buthavebeenerroneouslypredictedasfalse(N).

Usingthesevalues,wecancalculatetheaccuracyofthemodel. Theaccuracyisgivenby:



* + **Precision**:refersto the level of accuracyattainedin accuratepredictions,specificallyintermsofcorrectlylabeled observations.Itquantifiestheproportionof true observations that are correctly labeled as true. The formula to calculateprecisionisTP/(TP+FP).
  + **Recall**, also referred to as Sensitivity, is a metric that quantifies the number ofcorrectlypredictedobservationsof thetrue class.Itiscalculatedbydividingthe

numberof truepositives (TP) bythesumof truepositives andfalse negatives (TP

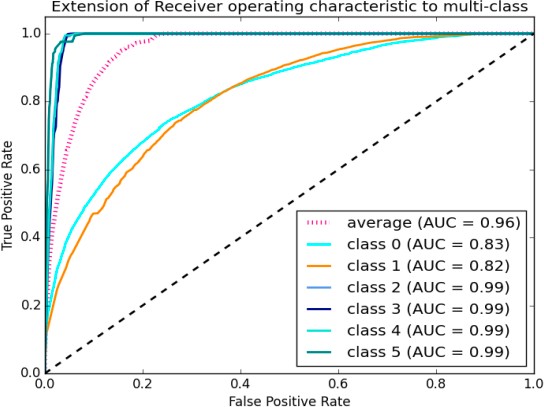
+FN).

* + **Specificity** referstotheaccuracy ofcorrectly identifyingobservationsofthefalseclass. It is calculated by dividing the true negatives (TN) by the sum of truenegatives andfalsepositives(TN+FP).

Specificityand Sensitivityplaysa crucial roleinderivingROCcurve.

* The ROC curve is a graphical representation that summarizes the performance of amodelbyanalyzingthebalancebetween thetrue positiverate (sensitivity)and thefalsepositiverate(1-specificity).
* The area underthecurve(AUC),also knownas the indexofaccuracy (A)orconcordance index, serves as an excellent performance metric for the ROC curve. AhigherAUC indicatesastrongerpredictivepowerofthe model.

ThisishowaROCcurvelookslike:



#### Figure6.2:ROCcurve

The curve's area quantifies the model's ability to accurately classify true positives andtrue negatives. Our objective is for the model to correctly predict true classes as true andfalseclassesasfalse.

Hence, it can be stated that we aim for a true positive rate of 1. However, we are notsolelyconcernedwiththetruepositive rate,butalsowiththefalsepositiverate.Inourspecific problem, we not only want to predict the Y classes as Y, but also want the Nclasses tobepredictedasN.

Westrive tomaximizethearea underthecurve,whichishighestfor class2,3,4,and5in theaforementionedexample.

Forclass 1,when thefalse positive rate is0.2,thetrue positive rate isapproximately0.6.Conversely,forclass2,the true positiverateis 1 at thesame falsepositiverate.

Consequently,theAUCforclass2willbesignificantlyhighercomparedtotheAUCforclass1.Therefore,themodel forclass2willbesuperior.

Themodelsforclass2,3, 4,and5willpredictwithgreateraccuracy comparedtothemodels forclass0and1,astheAUC is higherforthose classes.

#### 7. RESULTANALYSIS

This project uses above mentioned models with both BOW and TF-IDF. Theaccuraciesofthemodelsarecompared.Themodelwhichgetsbestaccuracyistaken.

#### 8. CONCLUSION

Review classification using NLTK, Spacy, Scikit-learn is a machine learningmodel which uses modern Natural language processing techniques. With this Project, thepeople can easilyclassifythe review ofthemovie withoutputting much effort.Themodernalgorithmsinthisprojectplayskey roleingettinghighaccuracyfor agivendata.The MLP classifier gives the High accuracy for this project. Hence we can conclude thatMLP classifieristhe bestfitting modelforreviewClassification.

#### 9. FUTUREENHANCEMENTS

Textclassification involves assigningrelevantcategoriesfroma predefined settonatural language texts.Insimplerterms,it istheextractionof generictagsfromunstructuredtext. Thesegenerictagsarederivedfromapredeterminedsetofcategories.

## 10. REFERENCES

1. Liu,B.,"InvestigatingUserOpinionsinRecommenderSystems", ProceedingsoftheSecond KDD Workshop on Large Scale Recommender Systems and the Netflix PrizeCompetition,August 24,2008,LasVegas,Nevada,USA.
2. Dave, D., Lawrence, A., Pennock, D., "Analyzing the Peanut Gallery: ExtractingOpinions and Classifying Product Reviews", Proceedings of the International WorldWideWebConference(WWW‟03),2003.
3. Turney,P., "ThumbsUporThumbsDown?ApplyingSemanticOrientationtoUnsupervised ReviewClassification",ACL‟02,2002.
4. Porter, M., "AStemmingAlgorithmforTextClassification", Retrievedfrom<http://people.scs.carleton.ca/~armyunis/projects/KAPI/porter.pdf>.
5. AnIntroductiontoLogisticRegressionAnalysisandReporting,ArticleinTheJournalofEducationalResearch,September2002.
6. Mitchell,T.,"Decision TreesforTextClassification",Retrievedfrom<https://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf>.
7. Xu, B., "AnEnhancedRandomForestClassifierforTextCategorization",ShenzhenGraduateSchool,Harbin Institute of Technology,Shenzhen 518055,China.
8. Knigam,C.,"Multinomial LogisticRegressionforTextClassification",Retrievedfrom<http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf>.
9. Suykens, J.A.K.,Vandewalle, J.,"TrainingMultilayerPerceptronClassifiersUsingaModifiedSupportVectorMethod".