
Personal Care Monitoring System

Project report submitted in partial fulfillment of the requirement for
the degree of Bachelor of Technology

In

COMPUTER SCIENCE ENGINEERING

By:

Aayush Verma (171259)

under the supervision

of

Dr. Yugal Kumar

To




Department of Computer Science Engineering and Information Technology
Jaypee University of Information Technology, Wagnaghat, Solan,
Himachal Pradesh-173234

Candidate's Declaration

I hereby declare that the work presented in this report entitled “Personal Care Monitoring System” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from Jan 2021 to May 2021 under the supervision of **Dr.Yugal Kumar, Computer Science/ IT.**

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

A rectangular box containing a handwritten signature in blue ink that reads "Aayush".

Aayush Verma (171259)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

A handwritten signature in black ink that reads "Dr. Yugal Kumar".

Dr. Yugal Kumar

Computer Science/Information Technology

Dated:

Acknowledgement

Working for the “sentiment analysis system” project was interesting. We got to learn about NLP and some web techniques and to make usage more accessible and to make it more accurate.

Exceptional gratitude to our supervisor Dr. Yugal Kumar for his direction and guidance on this undertaking.

Additionally, we are exceptionally appreciative of our school, talks, and companions where they gave us sufficient opportunity to finish this report and we might want to show gratitude to every single other who bolstered us in our task.

Thankyou.

Table of Content

S. No.	Topic	Page No.
1.	Introduction	1
	Motivation	3
	Aims and Objectives	3
2.	Literature Survey	
3.	Feature Selection in Sentiment Classification	6
	Feature Selection Methods	7
	Search Based Feature Selection	10
4.	Sentiment Analysis Techniques	12
	Machine Learning Approach	14
	Lexicon Based Approach	18
5.	Lexicon Based Approach	21
6.	Naïve Bayes Classifier	38
7.	Technical Challenges	50
8.	Screenshots	53
9.	Conclusion	58
10.	Future Work	60
11.	References	62

List of Figures

S.No.	Title	Page No.
1.	Target of Sentiment Analysis	2
2.	Sentiment Analysis Techniques	13
3.	Support Vector Machines	16
4.	LBA I	22
5.	LBA II	25

List of Tables

S.No.	Title	Page No.
1.	Accuracy of different sentiment classifiers	20

Abstract

Sentiment Analysis (SA) is an ongoing field of studies within the textual content mining subject. SA is the computational treatment of reviews, sentiments and subjectivity of textual content.

The selection-making technique of people is laid low with the opinions shaped through thought leaders and ordinary humans. When someone needs to shop for a product on line he or she will be able to typically start by searching for evaluations and opinions written by different people at the various services. Sentiment analysis is one of the most up to date research areas in pc science. Over 7,000 articles were written on the topic. Hundreds of startups are growing sentiment analysis solutions and essential statistical applications which include SAS and SPSS consist of devoted sentiment analysis modules. There is a big explosion nowadays of 'sentiments' available from social media which includes Twitter, Facebook, message boards, blogs, and user forums. These snippets of text are a gold mine for businesses and individuals that want to screen their reputation and get well timed feedback approximately their products and moves. Sentiment evaluation gives these corporations the ability to screen the one of a kind social media web sites in real time and act as a consequence. Marketing managers, PR companies, campaign managers, politicians, and even fairness buyers and online buyers are the direct beneficiaries of sentiment analysis era.

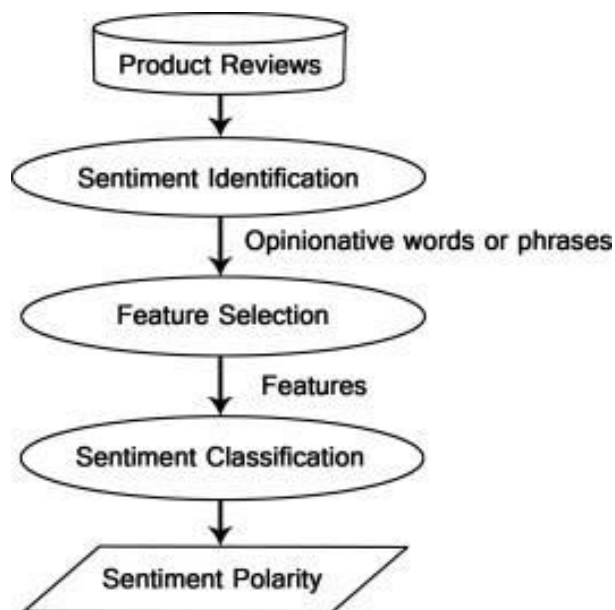
This record tackles a comprehensive assessment of the closing replace on this field and a complicated categorization of the strategies used in Sentiment Analysis.

The report additionally presents two techniques to extract records about the users' sentiment polarity (nice, impartial or negative), as transmitted within the messages they write. It additionally mentions the capability improvements that may be made to these method.

CHAPTER: 1

INTRODUCTION

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of humans's reviews, attitudes and emotions in the direction of an entity. The entity can represent individuals, activities or subjects. These subjects are maximum probably to be protected by way of critiques. The two expressions SA or OM are interchangeable. They specific a mutual that means. However, a few researchers said that OM and SA have slightly different notions. Opinion Mining extracts and analyzes humans's opinion about an entity even as Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to locate reviews, become aware of the emotions they specific, after which classify their polarity as proven in parent.



Sentiment Analysis can be considered a classification process as illustrated in figure. There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the

sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions. Sentiment expressions are not necessarily subjective in nature. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents. Classifying text at the document level or at the sentence level does not provide the necessary detail needed for opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity like this sentence *“The voice quality of this phone is not good, but the battery life is long”*.

1. Motivation

The social network sites and micro-blogging sites are considered a very good source of information because people share and discuss their opinions about certain topics freely. This project will be used to recognize people's emotions about those topics. In political debates for example, we could figure out people's opinions on certain election candidates or political parties. The election results can also be predicted from political posts.

Sentiment analysis can also be used in adaptive E-learning systems. In particular, affective and emotional factors, among other aspects, seem to affect the student motivation and, in general, the outcome of the learning process. Therefore, in learning contexts, being able to detect and manage information about the students' emotions at a certain time can contribute to know their potential needs at that time. On one hand, adaptive e-learning environments can make use of this information to fulfill those needs at runtime: they can provide the user with recommendations about activities to tackle or contents to interact with, adapted to his/her emotional state at that time. On the other hand, information about the student emotions towards a course can act as feedback for the teacher. This is especially useful for online courses, in which there is little (or none) face-to-face contact between students and teachers and, therefore, there are fewer opportunities for teachers to get feedback from the students.

Knowing the users' emotions is useful not only in the educational context but also in many others (e.g., marketing, politics, online shopping, and so on).

2. Aim and Objective

The aim of this project is to extract information about the users' sentiment as transmitted in the text they write on social networking sites, by assigning a score to each subsequent word in the text and computing the cumulative score

Chapter 2

Literature Survey

In this section, we will be summarizing the research papers we have referred to in order to conduct a comprehensive study of related to our project.

2.1 “Apps for Mental Health: An Evaluation of Behavior Change Strategies and Recommendations for Future Development”

Authors :FelwahAlqahtani, Ghazayil Al Khalifah, Rita Orji

Nowadays, mental health troubles have turn out to be a first-rate public health task. People with intellectual fitness problems locate it tough achieving their daily tasks which includes paintings and look at (Keyes, 2005). As result, lots of them are the usage of virtual programs to assist their intellectual fitness and beautify lifestyles pleasant. More than 10,000 intellectual health and wellness apps are to be had for down load and use (Torous and Roberts, 2017) on-line. The ubiquitous nature of smartphones and different hand held mobile devices are shaping-up users' lifestyles by using adding new aspects to the idea of socializing, accomplishing movements, and growing new habits (Oulasvirta et al., 2012). Therefore, smartphones are attractive platforms for researchers to deliver interventions. Mobile applications (apps) are being used to deliver interventions focused on various fitness issues (Iacoviello et al., 2017). For mental fitness troubles in particular, Roepke et al. (2015) and Areean et al. (2016) highlighted in their studies that mobile-based intellectual fitness intervention made a sturdy impact on lowering depressed mood. However, in addition they pronounced a excessive charge of drop-out.

By making use of diverse persuasive strategies to enhance, alternate, or form customers' behavior and/or attitudes, mental fitness apps can efficiently characteristic as guide tools that still motivate and stimulate customers to hold on

the use of the apps to attain higher mental health. However, the volume to which available mental fitness apps correctly employed persuasive strategies and how they put in force them in their app to achieve their supposed goal stays unknown.

Therefore, this paper pursues to gain three fundamental targets. First, we evaluate 103 intellectual fitness programs and perceive awesome persuasive techniques integrated in them using the Persuasive Systems Design (PSD) model and Behavior Change Techniques (BCTs). We in addition classify the persuasive strategies based totally at the kind of mental fitness problems the app is centered on. Second, we display the diverse methods that the persuasive strategies are implemented/operationalized in intellectual health applications to reap their supposed goals. Third, we look at whether or not there is relationship between apps effectiveness (measured through person ratings) and the persuasive strategies hired. To obtain this, two researchers independently downloaded and used all identified apps to discover the persuasive strategies the use of the PSD model and BCTs. Next, they also have a look at the diverse methods that those strategies had been applied within the intellectual health apps. The consequences display that the apps hired 26 awesome persuasive strategies and various 1–10 techniques in line with app. Self-tracking (n = fifty nine), personalization (n = fifty five), and reminder (n = forty nine) were the most regularly hired. We also observed that anxiety, pressure, melancholy, and popular mental health problems have been the not unusual mental health problems targeted by using the apps. Finally, we provide some layout tips for designing mental fitness apps primarily based on our consequences.

Identifying the persuasive techniques in intellectual health apps, classifying them based totally at the type of mental health problems the apps target, and uncovering the connection between app effectiveness and persuasive strategies hired could be treasured for each researchers and builders working in the mental fitness domain to inform the design of intellectual health apps.

2.2 Project SAM: Developing an app to provide self-help for anxiety

Authors :

Phil Topham, Department of Health and Social Sciences

Praminda Caleb-Solly, Department of Computer Science and Creative Technologies

Paul Matthews, Department of Computer Science and Creative Technologies

University of the West of England, Bristol

October 2015

An interdisciplinary team on the University of the West of England (UWE) became commissioned and funded to expand a cell phone app which could offer self-assist alternatives for the control of mild to slight tension. The finished app would expand the variety and availability of psychological aid for pupil well-being at UWE and other higher education establishments.

The project crew consisted of computer scientists and one psychologist who have been responsible for the technical, practical and scientific specification of the app. A nearby mobile app development company changed into appointed and the groups collaborated at the design, build and assessment of the app.

The self-assist structure and components have been advanced in session with therapeutic

practitioners, in and out of UWE. The developer team cautioned on and built multi-media features to recognize the self-assist goals of the app.

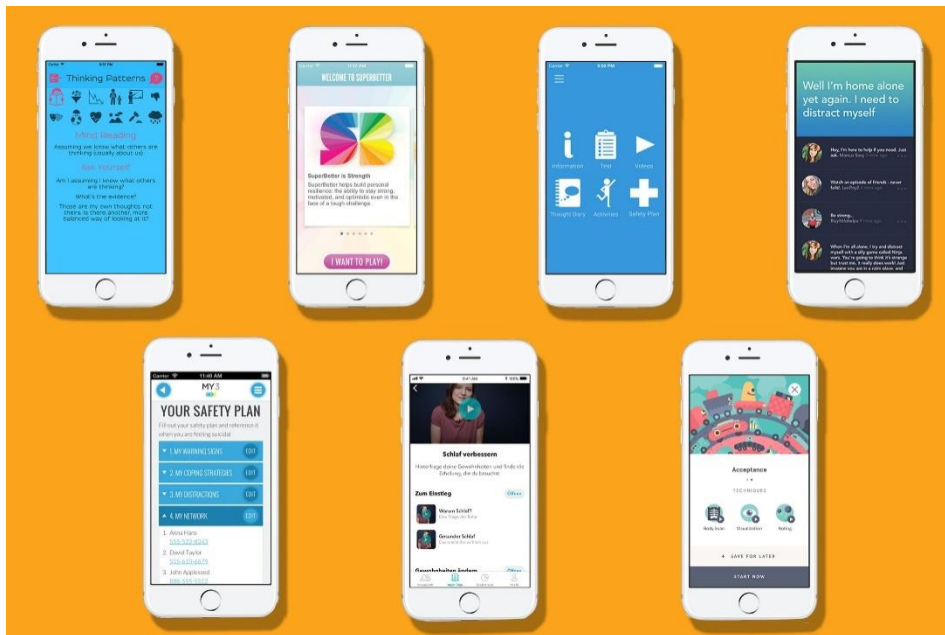
The UWE mission group promoted an iterative approach to improvement, comparing every level of development via trials with expert users, practitioners and college students. The app, named SAM (Self-assist for Anxiety Management), changed into evolved for Apple and Android working systems, to be usable on smartphones and tablets.

SAM turned into released within the app shops in July 2013, globally to be had and unfastened to download for the first 12 months of operation. It was promoted to college students, educational establishments, intellectual

health organisations and charities as well as quite a number professional and casual contacts.

A UWE-based totally Advisory Board changed into convened to supervise the protection and improvement of the university's funding in SAM. Members consist of the project team, researchers, therapists and other staff with an interest in its use to help pupil properly-being.

Three key responsibilities of the Board are to make certain SAM's economic sustainability, to supervise traits in its usability and self-help components, and to reap investment for the evaluation of its healing impa.



2.3 “Development of a Mobile Phone App to Support Self-Monitoring of Emotional Well-Being: A Mental Health Digital Innovation”

Authors:

Nikki Rickard BBSc (Hons), PhD (Emotion and Well-being Research Unit, School of Psychological Sciences, Monash University, Clayton, Australia ,

Centre for Positive Psychology, Melbourne Graduate School of Education, University of Melbourne, Melbourne, Australia)

Hussain-AbdulahArjmand, BBNSc (Hons) (Emotion and Well-being Research Unit, School of Psychological Sciences, Monash University, Clayton, Australia)

Emotional well-being is widely defined as, “a high-quality sense of properly-being and an underlying notion in our personal and others’ dignity and really worth” by the Mental Health Foundation (p. 8). Consistent with twin models of properly-being, it encompasses both fine functioning (happiness, a feel of manipulate and self-efficacy, and social connectedness) and an absence of pressure and despair. Monitoring adjustments in emotional properly-being is fundamental to mental fitness, with increases in emotional well-being associated with resilience, creative questioning, social connectivity, and bodily health. In contrast, massive and sustained decreases in emotional well-being are associated with the development of affective issues consisting of melancholy and anxiety, and reduced physical fitness.

Monitoring for such modifications is vital for early detection of intellectual health problems. Rapid reaction to early risk indicators is one of the key predictors of higher fitness effects, permitting preventative fitness processes to be initiated early. Regular tracking of emotional health indices is consequently advocated with the aid of numerous national suggestions. In exercise, however, it remains tough for clinicians or professional mental fitness carrier carriers to achieve frequent monitoring in actual time. A precedence project going through the fitness care device is to obtain conceivable and sustainable means of assisting self-control of

health and well-being. Self-tracking is a specially attractive goal for mental health care, for the reason that many individuals with intellectual fitness needs do not are looking for expert fitness care aid. In addition, self-monitoring may also broaden an person's insight into their need to searching for assist. In specific, young human beings constantly suggest that they select nonprofessional or self-controlled techniques for addressing intellectual fitness troubles. Obtaining temporally touchy (eg, day by day) facts on tremendous modifications in emotional nation has the capability to profoundly improve the capacity to promote emotional fitness.

Experience sampling methodologies (ESMs), or ecological temporary assessments, involve the systematic collection of self-report statistics from individuals at multiple time factors for the duration of their everyday lives. ESMs were used to display changes in affective state, and to are expecting mental fitness with achievement to a certain quantity. In specific, the variety in emotional state over the years gives greater full-size facts for information the reasons and nature of psychopathology than do go-sectional "snapshot" exams. For instance, whilst sampled more than one times a day for 6 days, negative affect changed into located to vary greater in sufferers identified with most important depressive ailment than that in controls throughout the day. ESM checks in people recognized with panic sickness also found out that the expectancy of a panic attack become a massive precursor for the incidence of a panic assault. Ben-Zeev et al also found that sufferers recognized with a major depressive ailment retrospectively reported better tiers of signs and symptoms regarding anhedonia, suicidality, and unhappiness than captured in their ESM reports, highlighting the biases of conventional survey methods. To date, but, it has been methodologically tough and obvious to gain temporally ordinary and specific measures of emotional state. The assets required to attain such information again and again over prolonged time frames have made such an extensive monitoring prohibitive. In addition, the use of palm pilots and pagers (which were by no means as acquainted to customers as cell phones have come to be) to set off users for this statistics may be intrusive, and makes it much less possibly that users will retain to use this form of tracking for extended periods.

Mobile cellphone era offers an exceptional opportunity to unobtrusively song

ordinary behavior and adjustments in emotional nation, all in actual time. Mobile smartphone fitness tools additionally offer the potential of immediately response to the outcome of this tracking through shipping of mental health information contingent on changes in real-time emotional country. This era has not yet been completely leveraged for these purposes, regardless of cell telephones being one of the few portions of generation that the general public keep on their man or woman each day. This pervasiveness method that mobile telephones provide a exceptionally herbal and normal approach with the aid of which facts on emotional country might be acquired. Mobile phones now penetrate seventy seven%, 72%, and sixty eight% of the Australian, US, and UK population, respectively, and are a price-powerful approach of seeking assist for intellectual fitness troubles which can overcome socioeconomic and geographic obstacles.

Mobile phone health technology holds fantastic ability for facilitating the control of emotional health through its potential to supply bendy, user-orientated intervention and self-management equipment; a feature especially applicable for younger those who often file fear of stigma related to seeking professional services for touchy intellectual fitness troubles. In a 2010 look at, seventy six% of an Australian sample reported being inquisitive about the usage of mobile telephones to monitor and manage their own intellectual fitness. A large quantity of cellular telephone apps are currently available that claim to sell intellectual health and properly-being and a subset of these also try to song temper or emotional country over the years. However, empirical aid for the efficacy of those apps is extraordinarily restricted. For example, in a scientific overview of 5464 mental fitness app abstracts, less than 5 apps had been observed to have experimental proof. In addition, a few have capitalized at the blessings enabled by the cellular telephone generation consisting of enjoy sampling and automated records collection in identifying and comparing capacity time-touchy behavioral signs of mental fitness change.

2.4 “Utilization of Patient-Generated Data Collected Through Mobile Devices: Insights From a Survey on Attitudes Toward Mobile Self-Monitoring and Self-Management Apps for Depression”

Authors:

Ralf Hartmann, MSc(Research Center of the German Depression Foundation, Leipzig, Germany)

Christian Sander, Dr(Research Center of the German Depression Foundation, Leipzig, Germany)

Noah Lorenz, MSc(Research Center of the German Depression Foundation, Leipzig, Germany)

Daniel Böttger, MA(Research Center of the German Depression Foundation, Leipzig, Germany)

Ulrich Hegerl, Prof Dr(Department of Psychiatry, University of Leipzig, Leipzig, Germany)

Depression is a extreme ailment with large consequences on well-being and high-quality of life. Major depressive disorder (MDD) is incredibly established and is a prime motive for years lived with disability and a first-rate source of the worldwide burden of ailment. Furthermore, MDD is related to a high hazard of recurrence and chronicity. Although diagnostics and evidence-based totally treatments (eg, pharmacotherapy, psychotherapy) for depressive problems are to be had, a sizeable percentage of people with a diagnosis of unipolar depressive disorder do now not receive remedy as recommended by national hints. Consequently, self-monitoring and self-management grow to be ever more essential. The opportunities that have arisen from the digital and cellular revolution of new years endure the capability to satisfy these challenges. Mobile devices which include smartphones or wearable biosensors can check and document multimodal data including physiological statistics, self-ratings, person behavior, or environmental records. Such affected person-generated facts become increasingly to be had and feature promising ability to be utilized for self-

tracking, self-management, and hospital treatment. However, this discipline of studies is young and it presentations a whole lot dynamic. Throughout any technique of development or implementation of mobile structures or apps for self-tracking and self-management, a profound knowledge of readiness for use and person behavior is necessary as common adherence to mobile fitness (mHealth) structures or apps is regularly weak. So a long way, only a few studies have explored alternatives and utilization of mHealth apps in popular or for unique fields of hobby but no longer for despair. The aim of this survey became to offer descriptive statistics to answer the subsequent questions: to what quantity are cellular apps for the self-tracking and self-management of melancholy (for the cause of clarity, the subsequent abbreviation is used henceforth: MSSD) currently used, what is the predicted destiny use, and what do capability customers opt for in terms of documented parameters and statistics-sharing alternatives

CHAPTER 3:

FEATURE SELECTION IN SENTIMENT CLASSIFICATION

Sentiment Analysis project is taken into consideration a sentiment classification hassle. Feature selection is frequently incorporated because the first step in gadget studying algorithms like SVM, Neural Networks, ok-Nearest Neighbors, and so forth. The important purpose of the function selection is to lower the dimensionality of the function space and as a consequence computational cost. As a 2d goal, characteristic choice will lessen the overfitting of the studying scheme to the schooling statistics. During this process, it is also vital to find a suitable tradeoff between the richness of capabilities and the computational constraints involved when fixing the categorization project. Some of the cutting-edge capabilities are:

Terms presence and frequency: These capabilities are individual words or word n-grams and their frequency counts. It both gives the phrases binary weighting (0 if the phrase seems or one if in any other case) or makes use of term frequency weights to indicate the relative importance of functions.

Parts of speech (POS): finding adjectives, as they're essential indicators of reviews.

Opinion phrases and phrases: those are phrases typically used to express evaluations consisting of desirable or bad, like or hate. On the alternative hand, a few terms express opinions with out the usage of opinion words. For example: fee me an arm and a leg.

Negations: the appearance of terrible phrases may additionally trade the opinion orientation like no longer desirable is equivalent to awful.

1. Feature selection methods

Feature Selection strategies may be divided into lexicon-primarily based strategies that want human annotation, and statistical strategies which are automated strategies which might be more regularly used. Lexicon-based totally tactics typically start with a small set of ‘seed’ words. Then they bootstrap this set through synonym detection or on-line assets to gain a bigger lexicon. This proved to have many problems as reported by means of Whitelaw et al. Statistical techniques, on the other hand, are fully computerized.

The feature selection strategies deal with the documents both as institution of phrases (Bag of Words (BOWs)), or as a string which keeps the collection of phrases within the record. BOW is used more frequently due to its simplicity for the type process. The most commonplace characteristic selection step is the removal of forestall-phrases and stemming (returning the word to its stem or root i.E. Flies → fly). In the following subsections, we gift three of the maximum frequently used statistical strategies in FS and their associated articles. There are other strategies used in FS like statistics advantage and Gini index.

1.1. Point-wise Mutual Information (PMI)

The mutual data measure gives a formal way to version the mutual records among the features and the instructions. This measure was derived from the data principle. The factor-wise mutual data (PMI) $M_i(w)$ among the phrase w and the elegance i is defined on the premise of the extent of co-incidence among the class i and word w . The expected co-occurrence of class i and phrase w , on the idea

of mutual independence, is given by $P_i \cdot F(w)$, and the proper co-incidence is given

by way of $F(w) \cdot p_i(w)$.

The mutual statistics is described in phrases of the ratio among those two values and is given by using the subsequent equation:

$$M_i(w) = \log \left(\frac{F(w) \cdot p_i(w)}{F(w) \cdot P_i} \right) = \log \left(\frac{p_i(w)}{P_i} \right)$$

The word w is positively correlated to the class i , when $M_i(w)$ is greater than 0. The word w is negatively correlated to the class i when $M_i(w)$ is less than 0.

1.2. Chi-square (χ^2)

Let n be the overall quantity of files in the series, $p_i(w)$ be the conditional chance of class i for files which contain w , P_i be the worldwide fraction of files containing the class i , and $F(w)$ be the worldwide fraction of documents which contain the phrase w . Therefore, the χ^2 -statistic of the phrase between word and class i is defined as

$$\chi_i^2 = \frac{n \cdot F(w)^2 \cdot (p_i(w) - P_i)^2}{F(w) \cdot (1 - F(w)) \cdot P_i \cdot (1 - P_i)}$$

1.3. Latent Semantic Indexing (LSI)

Feature choice methods try to reduce the dimensionality of the records by picking from the original set of attributes. Feature transformation methods create a smaller set of functions as a function of the original set of functions.

In the following subsections, we present 3 of the most regularly used statistical techniques in FS and their associated articles. There are other methods utilized in FS like data gain and Gini index.

1.4. Point-wise Mutual Information (PMI)

The mutual facts degree provides a formal manner to version the mutual statistics among the functions and the training. This degree was derived from the information principle. The factor-sensible mutual data (PMI) $M_i(w)$ among the phrase w and the elegance i is described on the idea of the extent of co-incidence between the magnificence i and word w . The anticipated co-incidence of sophistication i and word w , on the idea of mutual independence, is given through $P_i \cdot F(w)$, and the real co-prevalence is given

by means of $F(w) \cdot p_i(w)$.

The mutual facts is defined in terms of the ratio among those two values and is given by way of the following equation:

$$M_i(w) = \log \left(\frac{F(w) \cdot p_i(w)}{F(w) \cdot P_i} \right) = \log \left(\frac{p_i(w)}{P_i} \right)$$

1.5. Chi-square (χ^2)

Let n be the entire variety of files inside the collection, $p_i(w)$ be the conditional opportunity of class i for documents which comprise w , P_i be the global fraction of documents containing the class i , and $F(w)$ be the worldwide fraction of documents which include the phrase w . Therefore, the χ^2 -statistic of the phrase among word and magnitude is defined as

$$\chi_i^2 = \frac{n \cdot F(w)^2 \cdot (p_i(w) - P_i)^2}{F(w) \cdot (1 - F(w)) \cdot P_i \cdot (1 - P_i)}$$

1.6. Latent Semantic Indexing (LSI)

Feature selection techniques attempt to lessen the dimensionality of the data with the aid of picking from the authentic set of attributes. Feature transformation strategies create a smaller set of features as a characteristic of the authentic set of capabilities. LSI is one of the famous function transformation techniques. LSI method transforms the textual content area to a new axis

machine which is a linear combination of the authentic phrase capabilities. Principal Component Analysis techniques (PCA) are used to achieve this purpose. It determines the axis-system which keeps the greatest stage of records about the versions in the underlying characteristic values. The foremost downside of LSI is that it is an unmonitored technique which is ignorant of the underlying elegance-distribution. Therefore, the capabilities located by means of LSI are not necessarily the guidelines along which the class-distribution of the underlying documents may be excellent separated

It is usually stated that the ability to work with textual content on a semantic foundation is vital to modern facts retrieval structures. As a end

result, using LSI has considerably multiplied in recent years as earlier demanding situations in scalability and overall performance had been overcome.

2. Search based feature selection

An advantage of search based totally function choice strategies over ratings are commonly greater correct effects. These strategies are primarily based on both stochastic and heuristic searching techniques, what implies better computational complexity, which for very huge datasets that have some thousands of variables may limit some algorithms usability.

Typical solutions of search based totally feature selection are forward/backward choice methods.

Forward choice

Forward choice starts from an empty function set and, in every generation, adds one new attribute, shape the set of last. One that is introduced is this selection which maximizes positive criterion commonly classification accuracy. To make certain the right outcome of including a new feature to the function subset, first-rate is measured within the move validation system.

Backward elimination

Backward removal algorithm differs from forward choice with the aid of beginning from the full feature set, and iteratively removes one after the other function. In each new release simplest one characteristic is removed, which mostly influences ordinary version accuracy, as long as the accuracy stops increasing.

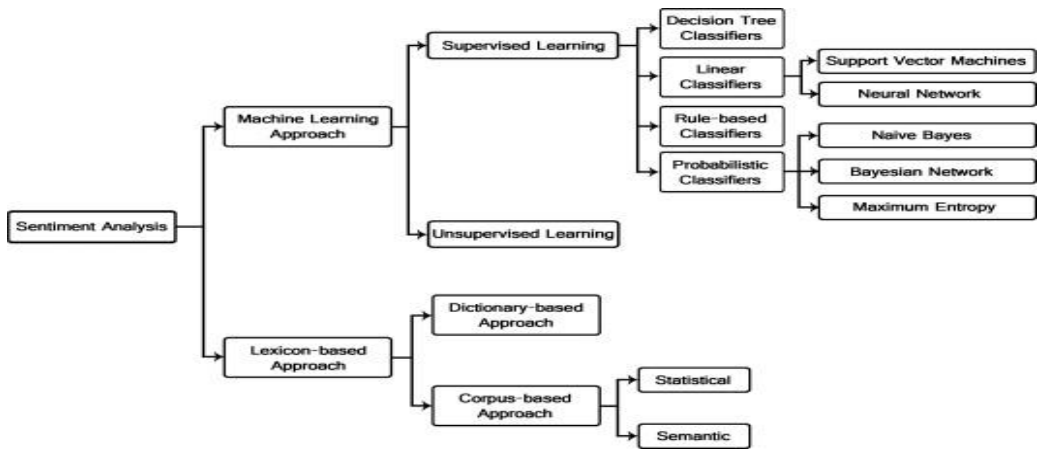
CHAPTER 4:

SENTIMENT ANALYSIS TECHNIQUES

Sentiment Classification strategies can be kind of divided into machine learning approach, lexicon based technique and hybrid approach. The Machine Learning Approach (ML) applies the well-known ML algorithms and uses linguistic capabilities. The Lexicon-based Approach relies on a sentiment lexicon, a collection of recognised and precompiled sentiment terms. It is split into dictionary-based method and corpus-based approach which use statistical or semantic techniques to find sentiment polarity. The hybrid Approach combines both methods and is very not unusual with sentiment lexicons playing a key position in most people of methods.

The text category methods using ML approach may be more or less divided into supervised and unsupervised studying methods. The supervised strategies make use of a large range of labeled training files. The unsupervised methods are used whilst it's far hard to find those categorized training documents.

The lexicon-primarily based approach relies upon on finding the opinion lexicon which is used to analyze the textual content. There are strategies on this technique. The dictionary-primarily based technique which relies upon on finding opinion seed words, and then searches the dictionary in their synonyms and antonyms. The corpus-based totally method starts offevolved with a seed list of opinion words, and then unearths different opinion words in a large corpus to help in finding opinion phrases with context specific orientations. This will be finished via using statistical or semantic strategies. There is a short clarification of each approaches' algorithms and related articles inside the subsequent subsections.



1. Machine learning approach

Machine learning approach relies at the well-known ML algorithms to solve the SA as a ordinary text class trouble that makes use of syntactic and/or linguistic functions.

Text Classification Problem Definition: We have a set of schooling facts $D = X_1, X_2, \dots, X_n$ in which every report is classified to a category. The class model is associated with the capabilities within the underlying document to one of the magnificence labels. Then for a given example of unknown class, the model is used to are expecting a class label for it. The hard class hassle is when simplest one label is assigned to an example. The tender classification problem is whilst a probabilistic fee of labels is assigned to an instance.

1.1. Supervised mastering

The supervised studying techniques depend on the lifestyles of classified education files. There are many varieties of supervised classifiers in literature. In the next subsections, we present in quick info some of the most frequently used classifiers in SA.

Probabilistic classifiers

Probabilistic classifiers use aggregate fashions for classification. The combination version assumes that each magnificence is part of the combination. Each mixture component is a generative version that offers the probability of sampling a selected time period for that factor. These sorts of classifiers are also known as generative classifiers. Three of the maximum well-known probabilistic classifiers are mentioned within the subsequent subsections.

- Naïve Bayes Classifier (NB): The Naïve Bayes classifier is the handiest and maximum normally used classifier. Naïve Bayes class model computes the posterior chance of a class, primarily based on the distribution of the phrases within the report. The version works with the BOWs function extraction which ignores the location of the phrase inside the report. It makes use of Bayes Theorem to expect the chance that a given

characteristic set belongs to a selected label.

$P(\text{label})$ is the prior opportunity of a label or the chance that a random characteristic set the label. $P(\text{features}|\text{previous})$ is the chance that a given function set is being categorized as a label. $P(\text{features})$ is the prior possibility that a given feature set is occurred. Given the Naïve assumption which states that each one features are impartial, the equation might be rewritten as follows:

- **Bayesian Network (BN):** The principal assumption of the NB classifier is the independence of the features. The other intense assumption is to assume that every one the capabilities are completely based. This results in the Bayesian Network version which is a directed acyclic graph whose nodes constitute random variables, and edges constitute conditional dependencies. BN is taken into consideration a complete version for the variables and their relationships. Therefore, a complete joint opportunity distribution (JPD) over all the variables is designated for a model. In Text mining, the computation complexity of BN is very high priced; that is why, it isn't always often used

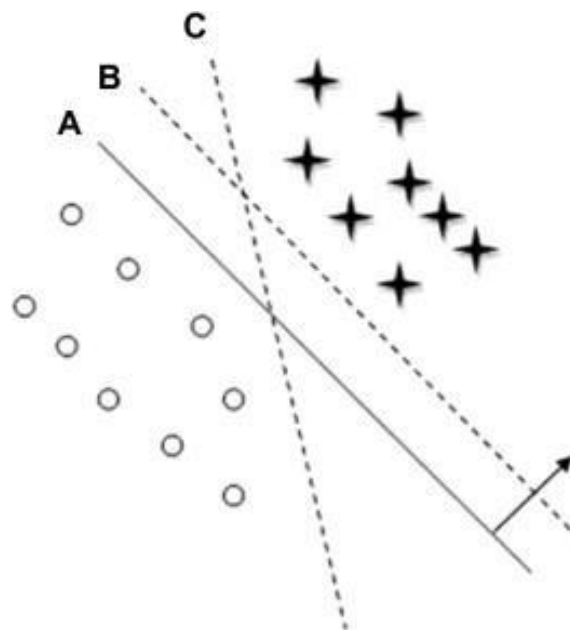
- **Maximum Entropy Classifier (ME):** The Maxent Classifier (called a conditional exponential classifier) converts classified function sets to vectors the usage of encoding. This encoded vector is then used to calculate weights for every function that may then be combined to decide the most in all likelihood label for a characteristic set. This classifier is parameterized through a hard and fast of X weights, which is used to mix the joint capabilities which can be generated from a feature-set by $\text{an}X$ encoding. In particular, the encoding maps each $C(\text{featureset}, \text{label})$ pair to a vector.

Linear Classifiers

Given x is the normalized record word frequency, w is a vector of linear coefficients with the identical dimensionality as the characteristic space, and b is a scalar; the output of the linear predictor is defined as $p = w \cdot x + b$, that's the output of the linear classifier. The predictor p is a keeping apart hyperplane between specific lessons. There are many kinds of linear

classifiers; amongst them is Support Vector Machines (SVM) that's a shape of classifiers that try and determine good linear separators among extraordinary classes. Two of the most famous linear classifiers are mentioned within the following subsections.

- **Support Vector Machines:** The primary principle of SVMs is to determine linear separators in the seek area that could first-class separate the special training. In the subsequent discern, there are 2 training x, o and there are 3 hyperplanes A, B and C. Hyperplane A offers the nice separation between the training, because the regular distance of any of the facts factors is the biggest, so it represents the most margin of separation. Text information are ideally suited for SVM type because of the sparse nature of textual content, in which few features are beside the point, however they tend to be correlated with each other and usually organized into linearly separable classes. SVM can assemble a nonlinear selection surface within the original function space through mapping the records times non-linearly to an inner product space where the classes can be separated linearly with a hyperplane.



SVMs are used in many applications, among these applications are classifying reviews according to their quality.

- **Neural Network (NN):** Neural Network consists of many neurons wherein

the neuron is its simple unit. The inputs to the neurons are denoted by means of the vector $\overline{X_i}$ that is the phrase frequencies inside the i th report. There

are a set of weights A which might be related to each neuron used for you to compute a feature of its inputs $f(\cdot)$. The linear feature of the neural network

is: . In a binary category problem, it is assumed that the magnificence label of is denoted through y_i and the signal of the expected function p_i yields the magnificence label. Multilayer neural networks are used for non-linear limitations. These multiple layers are used to set off a couple of piece-wise linear barriers, that are used to approximate enclosed regions belonging to a specific class. The outputs of the neurons in the earlier layers feed into the neurons inside the later layers. The schooling technique is more complicated due to the fact the mistakes want to be back-propagated over exceptional layers.

Decision Tree Classifiers

Decision tree classifier gives a hierarchical decomposition of the training records area in which a condition at the characteristic fee is used to divide the records. The circumstance or predicate is the presence or absence of 1 or extra words. The division of the data area is carried out recursively until the leaf nodes comprise positive minimal numbers of facts which are used for the motive of class.

There are different varieties of predicates which depend on the similarity of documents to correlate sets of terms which may be used to similarly partitioning of files. The extraordinary kinds of splits are Single Attribute cut up which use the presence or absence of precise words or phrases at a selected node in the tree with a purpose to perform the cut up. Similarity-based totally multi-characteristic cut up uses documents or frequent phrases clusters and the similarity of the documents to those phrases clusters on the way to perform the break up. Discriminant-based totally multi-characteristic split makes use of discriminants inclusive of the Fisher discriminate for performing the split.

Rule Based Classifiers

In rule based totally classifiers, the information space is modeled with a hard and fast of regulations. The left hand aspect represents a condition on the feature set expressed in disjunctive ordinary form even as the right hand side is the class label. The situations are at the time period presence. Term absence is hardly ever used because it isn't informative in sparse statistics.

There are numbers of standards in order to generate guidelines, the education phase construct all the rules depending on these criteria. The maximum two common standards are help and self belief. The assist is the absolute number of times within the schooling data set which can be relevant to the rule. The Confidence refers to the conditional possibility that the right hand facet of the rule is glad if the left-hand facet is happy.

1.2 Weakly, semi and unsupervised learning

The main purpose of text classification is to classify documents into a certain number of predefined categories. In order to accomplish that, large number of labeled training documents are used for supervised learning, as illustrated before. In text classification, it is sometimes difficult to create these labeled training documents, but it is easy to collect the unlabeled documents. The unsupervised learning methods overcome these difficulties.

2. Lexicon-based approach

Opinion words are hired in lots of sentiment type tasks. Positive opinion words are used to express some preferred states, even as negative opinion phrases are used to express some undesired states. There are also opinion terms and idioms which collectively are known as opinion lexicon. There are 3 important techniques which will compile or accumulate the opinion word listing. Manual method may be very time ingesting and it isn't used alone. It is generally blended with the other two computerized procedures as a very last check to keep away from the mistakes that resulted from computerized methods. The automated tactics are offered in the following subsections.

2.1. Dictionary-based totally method

A small set of opinion phrases is accumulated manually with recognised orientations. Then, this set is grown by means of looking within the widely

recognized corpora WordNet or glossary for their synonyms and antonyms. The newly located phrases are delivered to the seed listing then the next generation starts offevolved. The iterative process stops whilst no new words are discovered. After the manner is completed, manual inspection may be performed to eliminate or accurate errors.

The dictionary based totally method has a first-rate downside that is the incapability to find opinion phrases with area and context precise orientations.

2.2. **Corpus-primarily based technique**

The Corpus-based totally approach enables to solve the problem of locating opinion phrases with context specific orientations. Its strategies depend on syntactic styles or styles that occur collectively along with a seed list of opinion words to discover different opinion words in a massive corpus. We start with a list of seed opinion adjectives, and use them in conjunction with a

set of linguistic constraints to perceive extra adjective opinion phrases and their orientations. The constraints are for connectives like AND, OR, BUT, EITHER- OR.....; the conjunction AND as an instance says that conjoined adjectives generally have the same orientation. This concept is called sentiment consistency, which isn't always usually regular nearly. There are additionally adversative expressions such asbut, but which can be indicated as opinion adjustments. In order to decide if conjoined adjectives are of the equal or different orientations, learning is carried out to a huge corpus. Then, the links among adjectives form a graph and clustering is completed on the graph to provide two sets of words: nice and poor.

Using the corpus-based approach by myself is not as effective as the dictionary-primarily based technique due to the fact it's far hard to put together a large corpus to cover all English phrases, but this technique has a main gain that could assist to discover domain and context particular opinion phrases and their orientations the usage of a site corpus. The corpus-based totally method is executed the use of statistical method or semantic approach as illustrated inside the following subsections:

Statistical technique

Finding co-occurrence patterns or seed opinion words may be carried out

using statistical techniques. This might be performed via deriving posterior polarities the use of the co- incidence of adjectives in a corpus, as proposed with the aid of Fahrni and Klenner. It is viable to use the complete set of indexed files at the internet because the corpus for the dictionary construction. This overcomes the hassle of the unavailability of a few phrases if the used corpus is not large enough.

The polarity of a phrase can be identified via reading the prevalence frequency of the phrase in a big annotated corpus of texts. If the phrase takes place extra regularly amongst wonderful texts, then its polarity is effective. If it occurs extra frequently amongst poor texts, then its polarity is terrible. If it has same frequencies, then it's miles a neutral phrase. The comparable opinion phrases regularly seem collectively in a corpus. This is the principle remark that the kingdom of the art techniques are based on. Therefore, if two phrases appear collectively frequently within the identical context, they're likely to have the identical polarity. Therefore, the polarity of an unknown phrase can be decided by way of calculating the relative frequency of co-incidence with any other word. This can be finished using PMI.

Latent Semantic Analysis (LSA) is a statistical method which is used to analyze the relationships between a hard and fast of documents and the phrases noted in those documents to be able to produce a fixed of significant styles associated with the files and terms.

Semantic approach

The Semantic approach gives sentiment values at once and relies on one of a kind principles for computing the similarity among words. This principle offers comparable sentiment values to semantically close phrases. WordNet as an instance provides extraordinary kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for acquiring a listing of sentiment words by iteratively increasing the initial set with synonyms and antonyms and then figuring out the sentiment polarity for an unknown phrase by way of the relative count of tremendous and terrible

Accuracy of proposed classifiers.

Method	Accuracy (%)
Lexicon-based approach	80.02
Tree decision (J48-C4.5) + Lexicon-based tagging	83.17
Naïve-Bayes + Lexicon-based tagging	83.13
SVM + Lexicon-based tagging	83.27

synonyms

of

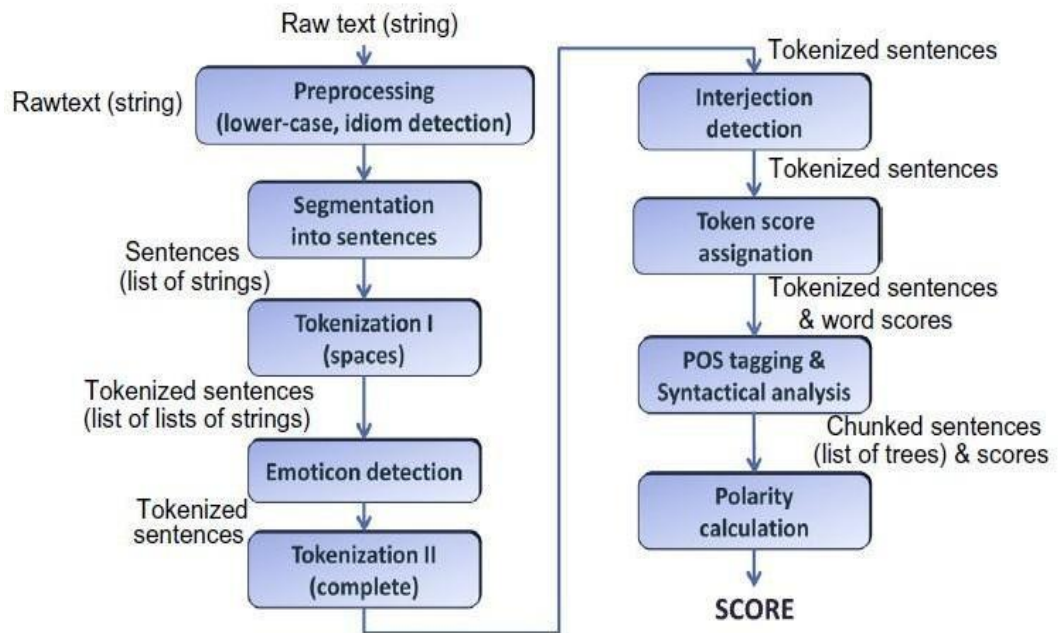
this

word.

CHAPTER 5:

LEXICON BASED APPROACH

Method I



Extracting sentiments from texts: message classification

1.1. Preprocessing (lower-case, idiom detection)

The first step consists of preprocessing the message to convert all the words into lower-case. Afterwards, it detects idioms and joins the words involved in each of them.

1.2. Segmentation into sentences

Then, the message is divided into sentences. Dots are the only punctuation marks considered as separators at this step, since others such as commas or semicolons can be part of emoticons.

1.3. Tokenization I (partial)

In the next step, tokens are extracted from each sentence. At this time, only

whitespaces are taken into consideration to separate tokens, since other separators, such as hyphens, can be part of emoticons.

1.4. Emoticon detection

Next, emoticons are detected. In order to detect them, the classifier in the text all the emoticons stored in two text files, containing positive and negative emoticons, respectively.

1.5. Tokenization II (complete)

During this second tokenization phase, all the punctuation marks (including commas, semicolons and so on) are considered as separators leading to obtain the final sets of tokens for each sentence.

1.6. Interjection detection

The next step consists of detecting and labeling interjections. Those that express laughs, such as “hehehe” or “hahaha”, are marked as positive whereas interjections such as “aaah” or “aaargh” are marked as negative. This is implemented through regular expressions, because in most of the cases, the interjections are intensified by repeating letters or sets of letters contained in the own word.

1.7. Token score assignation

The next phase consists of assigning a score to each token: 1 to 5 if it transmits a positive sentiment, 0 if it is neutral, and -1 to -5 if it is negative. To assign a score, the classifier checks if the token is a positive/negative emoticon, a positive/negative interjection, or whether it matches one of the words stored in the sentiment lexicon (L).

1.8. Removing repetitive letters

The next step is, for each token, to check whether it appears in any of the two dictionary categories and, if it is the case, to tag it as either positive or negative, accordingly. The messages written by users in Facebook usually contain very casual language. It is frequent to find words with repeated letters or with

non- alphabetic characters. Consequently, for each token, it if was not found in the dictionary in the form in which it appears in the message, letters occurring more than

twice in a row are replaced by only one occurrence and the new token is looked for in dictionary. Was this version of the token not found in the dictionary yet, then it is reduced to its lexeme.

1.9. Spelling checking

If the token does not match any word yet, then it is checked with a spelling checker. Since Facebook messages usually contain misspellings, a spelling checker is incorporated into the classifier. However, the spelling checker must be applied carefully, since some of the corrections it suggests produce bad results in the classifier. With the purpose of avoiding these situations, a list of words that should not be corrected, including names and surnames, is created and incorporated within the dictionary, so that, since they are found in the dictionary, they are not checked by the spelling checker. Finally, if a token is not classified into positive/ negative in any of the previous steps (even after all those considerations) the token is labeled as neutral.

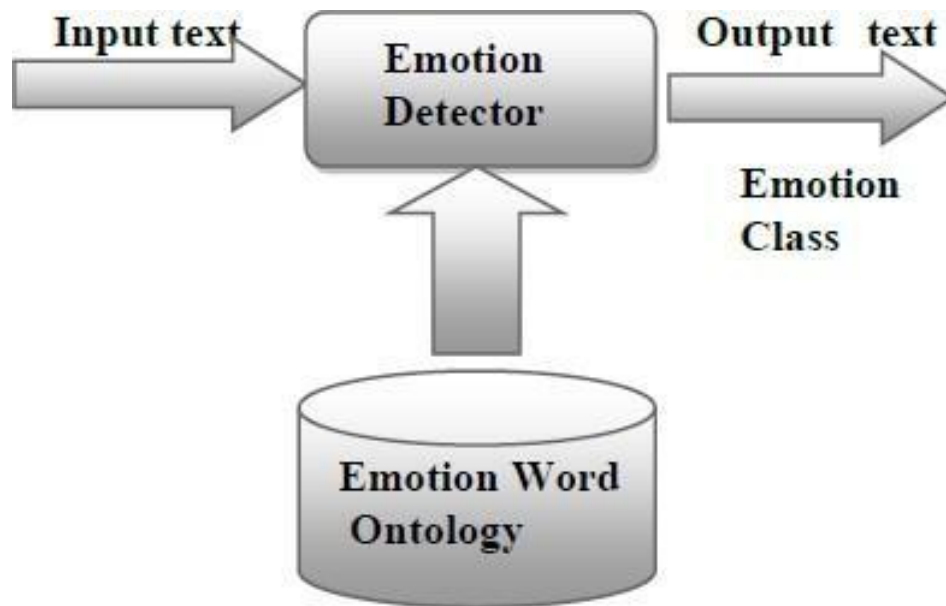
1.10. Syntactical analysis

Once each token has received a positive/neutral/negative score, each sentence is syntactically analyzed in order to check whether any score (positive/negative) should be reversed (e.g., because of negations). Firstly, we apply part of speech (POS) tagging to discriminate words that do not reflect any sentiment (e.g. articles) and to disambiguate words with multiple semantic meaning (e.g., words that can be both a noun and a verb). Afterwards, suffixes are analyzed. And then, negations are detected.

1.11. Polarity calculation

In order to calculate the polarity of a sentence, the number of tokens susceptible of conveying sentiments according to their grammatical category (i.e. noun, adjective, interjection or verb) is calculated. Other types of words, which appear frequently in texts (i.e. determinant, prepositions, etc.), are not considered, because they are “stopwords” for a sentiment analysis. Then, once each token has been scored, the final polarity score of a sentence is calculated.

2. Method II



The Framework is divided into two main components: Emotion Ontology, Emotion Detector.

2.1. Emotion Ontology

Ontology is an specific specification of conceptualization. Ontologies have definitional components like excessive degree schemas and elements like entities and attributes; interrelationship is between entities, area vocabulary. Ontologies offer an expertise of precise domain. Ontologies permit the domain to be communicated between men and women, institutions, and application systems. Emotion word hierarchy is converted into ontology. An ontology improvement device is used to broaden emotion ontology. The ontology has elegance and subclass courting format. Emotion training on the number one level in emotion hierarchy are at the pinnacle of emotion ontology and emotion training on the tertiary stage are at the bottom of ontology. High weight age is assigned to the top level emotion classes and low to the decrease stage emotion training.

2.2. Emotion Detector Algorithm

Emotion of the textual records may be diagnosed with the help of this emotion detection set of rules. The algorithm calculates weight for precise

emotion via including weights

assigned at each level of hierarchy and also calculates same for its counter emotion, then compares the both scores and greater one is taken because the detected emotion.

2.Three. Parameters Used

Algorithm is to calculate weight age to be assigned to specific emotion phrases in order that they can be taken care of in line with it. Certain parameters are required for this cause. The first step is calculation of parameters. This venture is achieved with the assist of Jena library which permits traversal and parsing of ontology.

Different parameters are calculated as follows:

Parent-Child relationship

If a text document belongs to a baby; it also indirectly refers back to the discern of it. Hence if a sure price is introduced to the child's score, discern rating additionally want to be changed. This is executed via traversing the ontology version in a breadth first way the use of Jena API. When any node is encountered all of its youngsters are retrieved. Then identical technique is carried out to every infant.

Depth in Ontology

This is required because it gives an idea approximately how specific is the term with regards to its corresponding ontology shape. The greater specific it's far the greater weight age ought to accept to it. This cost is calculated simultaneously even as traversing the ontology tree.

Frequency in Text file

This is also an critical parameter as extra is the frequency greater may be the importance of that term. This fee is calculated by means of parsing the text report and looking for occurrences of the phrases.

2.4 Algorithm

For each word in input text **i**

{ For each word in emotion lexicon **e**

{ Calculate length of **i**;

compare each letter of **i**with each letter of **e**;

if((no. of letters successfully compared=length of
i)&(next letter of e is an alphabet))
go to next input word;

```

        if((no. of letters successfully compared=length of
        i)|(next letter of e is '*'))
        { assign score of e to score
        of i; go to next input
        word;
        }
    }
}

```

3. Approaches to Building Sentiment Dictionaries

The computational speed and performance of dictionary-based totally procedures to sentiment analysis, together with their intuitive appeal, make such methods an appealing alternative for extracting emotional context from textual content. At the same time, each types of dictionary based processes provide capacity boundaries as nicely. Pre-built dictionaries to be used with modern preferred U.S. English have the gain of being extraordinarily smooth to use and significantly validated, making them sturdy contenders for programs in which the emotional content material of the language underneath look at is expressed in traditional approaches. At the identical time, the validity of such dictionaries rests severely on such conventional utilization of emotional words and terms. Conversely, custom dictionaries developed for unique contexts are touchy to versions in word usage, however come with a excessive fee of introduction and confined future applicability.

What we term specialised vocabularies arise in situations while the standard emotional valences related to unique words are no longer correct, either because words that typically convey emotional content do no longer do so inside the context in query or vice versa. For instance, in colloquial English the word “love” nearly continually carries a effective valence (and its inclusion in pre-built sentiment dictionaries reflects this fact) whilst the phrase “bagel” does now not. For professional and newbie tennis players, however, the two words would possibly mean some thing very unique; “love” method no factors scored (a scenario which has, if anything, a poor valence) and the

phrase “bagel” refers especially to the (terrible) occasion of losing a fixed 6-0 (e.G., “putting up a bagel within the first set”). It is simple to look how the software of a widespread sentiment dictionary to a body of textual content generated from a discussion of tennis should easily cause inaccurate inferences approximately its content material.

In such circumstances, a great approach is to expand a sentiment dictionary that reflects the emotional valence of the phrases as they may be utilized in that context. Such dictionaries mirror the emotional valence of the language as it's miles used in context, and so are much more likely to yield accurate estimates of sentiment in specialized vocabularies. Such dictionaries, but, are also tough and time-ingesting to construct, given that they typically involve specifying each emotionally-valenced phrase or phrase that might be encountered in that context. The project, then, is to expand an technique for constructing sentiment dictionaries inside the context of specialised vocabularies that is substantially greater efficient and less expensive than simple human coding.

CHAPTER 6:

NAIVE BAYES CLASSIFIER

Simple Bayesian classifiers were gaining recognition currently, and had been observed to carry out pretty properly. These probabilistic techniques make strong assumptions about how the records is generated, and posit a probabilistic version that embodies these assumptions; then they use a collection of categorized schooling examples to estimate the parameters of the generative version. Classification on new examples is achieved with Bayes' rule by way of deciding on the class this is maximum in all likelihood to have generated the instance. The naive Bayes classifier is the simplest of these fashions, in that it assumes that all attributes of the examples are unbiased of each different given the context of the class. This is the so-called "Naive Bayes assumption." While this assumption is in reality fake in maximum real-global responsibilities, naive Bayes regularly plays class very well. This paradox is defined via the fact that classification estimation is most effective a feature of the sign (in binary cases) of the feature estimation; the function approximation can nevertheless be bad even as class accuracy remains high. Because of the independence assumption, the parameters for every attribute can be found out separately, and this greatly simplifies mastering, especially while the variety of attributes is big. Document category is simply this sort of domain with a huge variety of attributes. The attributes of the examples to be labeled are phrases, and the wide variety of different words can be quite huge indeed. While some easy record class obligations may be appropriately finished with vocabulary sizes less than one hundred, many complex tasks on actual-global facts from the Web, UseNet and newswire articles do fine with vocabulary sizes inside the lots. Naive Bayes has been efficaciously applied to document type in many research efforts. Despite its popularity, there has been a few confusion within the report category community about the "Naive Bayes" classifier due to the fact there are two special generative fashions in not unusual use, both of which make the "Naive Bayes assumption." Both are called "Naive Bayes" by using their practitioners. One version specifies that a file is represented by a vector of

binary attributes indicating which phrases occur and do not occur in the document. The range of instances a phrase takes place in a file is not captured. When calculating the probability of a record, one multiplies the possibility of all of the characteristic values, consisting of the opportunity of non-occurrence for phrases that don't occur in the record. Here we will understand the document to be

the “occasion,” and the absence or presence of phrases to be attributes of the occasion. This describes a distribution based on a multi-variate Bernoulli occasion version. This approach is more conventional inside the field of Bayesian networks, and is appropriate for tasks that have a hard and fast quantity of attributes. The technique has been used for textual content type through several humans. The 2d version specifies that a file is represented through the set of phrase occurrences from the file. As above, the order of the phrases is misplaced; however, the range of occurrences of every phrase inside the document is captured. When calculating the probability of a document, one multiplies the opportunity of the words that occur. Here we can recognize the person phrase occurrences to be the “activities” and the file to be the collection of phrase events. We call this the multinomial occasion version. This approach is greater traditional in statistical language modeling for speech popularity, in which it would be known as a “unigram language model.” This method has additionally been used for textual content class via several people. Results suggest that the multi-variate Bernoulli model on occasion performs better than the multinomial at small vocabulary sizes. However, the multinomial normally outperforms the multi-variate Bernoulli at huge vocabulary sizes, and almost usually beats the multi-variate Bernoulli whilst vocabulary length is chosen optimally for both. While from time to time the distinction in overall performance isn't always exceptional, on average across information sets, the multinomial offers a 27% reduction in blunders over the multi-variate Bernoulli.

1. Probabilistic Framework of Naive Bayes

Consider the assignment of textual content type in a Bayesian learning framework. This technique assumes that the textual content data become generated by a parametric version, and uses training statistics to calculate Bayes-most fulfilling estimates of the version parameters. Then, ready with

these estimates, it classifies new test files the use of Bayes' rule to show the generative version around and calculate the posterior possibility that a class could have generated the test file in question. Classification then will become a easy remember of choosing the most possibly class.

Both eventualities count on that text documents are generated by using a mixture model parameterized by way of θ . The mixture model includes aggregate components of a file with a sum of total opportunity over all combination components: Each document has a class label. We count on that there may be a one-to-one correspondence between training and aggregate version components, and for this reason use c_j to suggest each the j th mixture component and the j th magnificence. In this putting, (supervised mastering from categorized education examples), the usually "hidden" indicator variables for a mixture version are furnished as those magnificence labels.

1.1 Multi-variate Bernoulli Model

In the multi-variate Bernoulli occasion model, a record is a binary vector over the distance of words. Given a vocabulary V , every measurement of the gap t , $t \in 1, \dots$ corresponds to word w_t from the vocabulary. Dimension t of the vector for

report d_i is written Bit, and is both zero or 1, indicating whether or not word w_t occurs as a minimum once within the report. With one of these file representation, we make the naive Bayes assumption: that the possibility of each word taking place in a report is unbiased of the occurrence of different phrases in a report. Then, the possibility of a record given its magnificence from Equation 1 is sincerely the manufactured from the opportunity of the characteristic values over all phrase attributes:

Thus given a generating component, a report can be visible as a set of multiple independent Bernoulli experiments, one for each word in the vocabulary, with the possibilities for every of those word events defined by using every component c_j ; θ). This is equivalent to viewing the distribution over documents as being described by a Bayesian network, where the

absence or presence of every phrase is dependent most effectively at the elegance of the document.

Given a hard and fast set of labeled education files, $D = \{D_i\}$, studying the parameters of a probabilistic classification model corresponds to estimating every

of those elegance-conditional word chances. The parameters of a combination issue are written $\theta_{wt|in}$ which $0 \leq \theta_{wt|gold}$ standard estimates for these chances by way of sincere counting of activities, supplemented via a previous. We use the Laplacean earlier, priming each phrase's depend with a count of 1 to keep away from probabilities of 0 $\in (0, 1)$ as given by the report's magnificence label. Then, we estimate the chance of phrase w_t in elegance c_j with:

The elegance prior parameters, θ_{c_j} , are set by using the maximum likelihood estimate:

Note that this version does no longer seize the quantity of times each phrase happens, and that it explicitly consists of the non-prevalence chance of phrases that do not seem inside the file

1.2 Multinomial Model

In evaluation to the multi-variate Bernoulli event model, the multinomial model captures word frequency statistics in files. Consider, as an instance, the prevalence of numbers within the Reuters newswire articles; our tokenization maps all strings of digits to a not unusual token. Since each information article is dated, and as a consequence has a number, the number token inside the multi-variate Bernoulli occasion model is uninformative. However, news articles about income generally tend to have a number of numbers in comparison to popular news articles. Thus, taking pictures frequency data of this token can help classification. In the multinomial model, a document is an ordered collection of phrase activities, drawn from the equal vocabulary V . We assume that the lengths of documents are independent of class. We again make a comparable Naive

Bayes assumption: that the possibility of each word event in a record is independent of the phrase's context and position inside the report. Thus, each file d_i is drawn from a multinomial distribution of phrases with as many independent trials because the period of d_i . This yields the familiar "bag of words" representation for files. Define N_{it} to be the total number of instances word w_t happens in record d_i . Then, the probability of a record given its class from Equation 1 is certainly the multinomial distribution:

2. Classification

Given estimates of those parameters calculated from the education files, classification can be finished on check documents via calculating the posterior chance of every magnificence given the evidence of the check document, and deciding on the magnificence with the highest opportunity. We formulate this through applying Bayes' rule:

The right hand aspect can be increased by first substituting the use of Equations 1 and

four. Then the growth of man or woman terms for this equation is depending on the occasion version used. Use Equations 2 and three for the multivariate Bernoulli event model. Use Equations five and six for the multinomial.

CHAPTER 7:

TECHNICAL CHALLENGES

- When inspecting the sentences misclassified through my mission, I can see that most of the class mistakes are related to the underlying ambiguity of the natural language. In the subsequent examples, I even have first supplied sentences that were incorrectly categorized as terrible. In the second case, there are sentences incorrectly classified as having a fantastic sentiment via the model. All those examples display the complexity of all herbal language and the want for developing language precise heuristics to better seize phraseological expressions, contrasting statements, sarcasm, and allusions made by the author

Pos sentence classify as neg.

1. “Longley has constructed a remarkably coherent, horrifically vivid snapshot of those turbulent days.”
2. “Romanek keeps the film constantly taut... reflecting the character's instability with a metaphorical visual style and an unnerving, heartbeat-like score.”
3. “Compelling revenge thriller, though somewhat weakened by a miscast leading lady.”

Neg sentence classify as pos.

1. “A mechanical action-comedy whose seeming purpose is to market the charismatic Jackie Chan to even younger audiences.”
2. “It's not so much a movie as a joint promotion for the national basketball association and teenaged rap and adolescent poster-boy lil' bow wow.”
3. “Director Tom Dey demonstrated a knack for mixing action and idiosyncratic humor in his charming 2000 debut shanghai noon, but showtime's uninspired send-up of tv cop show cliches mostly leaves him shooting blanks.”

- Another technical challenge is the inability to detect and rectify

textese, or SMS language, such as that used on Twitter.

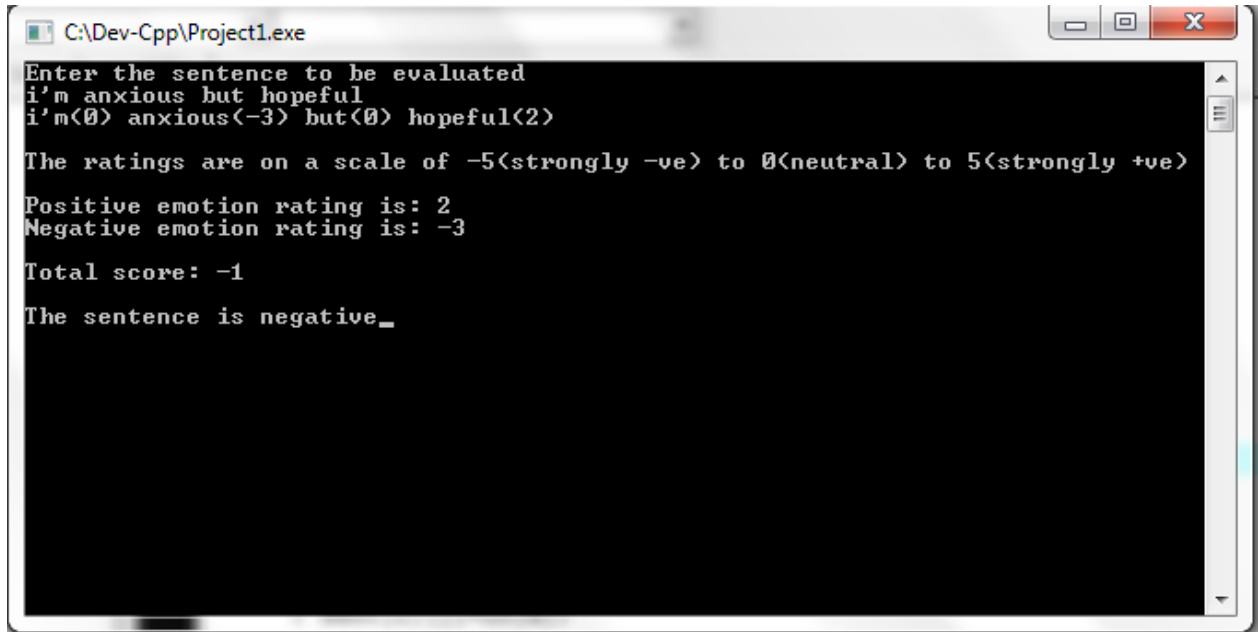
synonyms) could have different emotions in some extreme cases such as ironic or cynical sentences.

- Sentences without any keyword would imply that they do not contain any emotion at all, which is obviously wrong. For example, “I passed my qualify exam today” and “Hooray! I passed my qualify exam today” should imply the same emotion (joy), but the former without “hooray” could remain undetected if “hooray” is the only keyword to detect this emotion.
- Syntax structures and semantics also have influences on expressed emotions. For example, “I laughed at him” and “He laughed at me” would suggest different emotions from the first person’s perspective. As a result, ignoring linguistic information also poses a problem to keyword-based methods

CHAPTER 8:

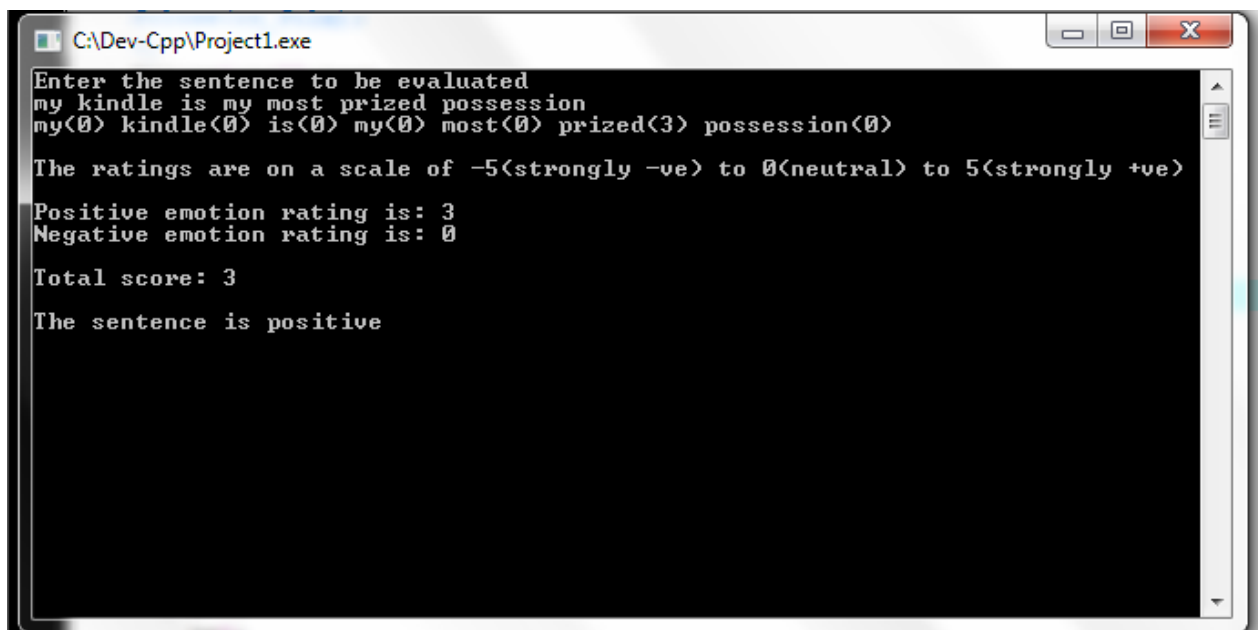
SCREENSHOTS

1. Lexicon Approach



```
C:\Dev-Cpp\Project1.exe
Enter the sentence to be evaluated
i'm anxious but hopeful
i'm(0) anxious(-3) but(0) hopeful(2)

The ratings are on a scale of -5(strongly -ve) to 0(neutral) to 5(strongly +ve)
Positive emotion rating is: 2
Negative emotion rating is: -3
Total score: -1
The sentence is negative_
```



```
C:\Dev-Cpp\Project1.exe
Enter the sentence to be evaluated
my kindle is my most prized possession
my(0) kindle(0) is(0) my(0) most(0) prized(3) possession(0)

The ratings are on a scale of -5(strongly -ve) to 0(neutral) to 5(strongly +ve)
Positive emotion rating is: 3
Negative emotion rating is: 0
Total score: 3
The sentence is positive
```

```
C:\Dev-Cpp\Project1.exe
Enter the sentence to be evaluated
alas! this was an absurd story
alas!(-2) this(0) was(0) an(0) absurd(-2) story(0)

The ratings are on a scale of -5(strongly -ve) to 0(neutral) to 5(strongly +ve)
Positive emotion rating is: 0
Negative emotion rating is: -4
Total score: -4
The sentence is negative
```

```
C:\Dev-Cpp\Project1.exe
Enter the sentence to be evaluated
he suffers from an incurable disease
he(0) suffers(-4) from(0) an(0) incurable(-2) disease(-3)

The ratings are on a scale of -5(strongly -ve) to 0(neutral) to 5(strongly +ve)
Positive emotion rating is: 0
Negative emotion rating is: -9
Total score: -9
The sentence is negative_
```

2. Machine Learning Approach

```
Problems @ Javadoc Declaration Console
<terminated> text [Java Application] C:\Program Files\Java\jdk1.8.0_05\bin\javaw.exe (May 13, 2015, 3:35:44 PM)
Usage: java MyLearner <fileData> <fileModel>
===== Loaded dataset: C:\Users\Vadehra\Desktop\try.txt =====

Correctly Classified Instances      145      85.7988 %
Incorrectly Classified Instances    24      14.2012 %
Kappa statistic                     0.6766
Mean absolute error                  0.1408
Root mean squared error              0.359
Relative absolute error              31.9983 %
Root relative squared error          76.599 %
Total Number of Instances           169

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          0.895   0.218   0.895     0.895   0.895     0.939   positive
          0.782   0.105   0.782     0.782   0.782     0.939   negative
Weighted Avg.  0.858   0.181   0.858     0.858   0.858     0.939

===== Evaluating on filtered (training) dataset done =====
===== Training on filtered (training) dataset done =====
===== Saved model: classy.dat =====
```

The model used for training the dataset has given an accuracy of 85% on 169 instances.

```
Problems @ Javadoc Declaration Console
<terminated> text2 [Java Application] C:\Program Files\Java\jdk1.8.0_05\bin\javaw.exe (May 13, 2015, 3:50:25 PM)
===== Loaded text data: C:\Users\Vadehra\Desktop\smstest.txt =====
I agree with Capital punishment today we find criminality is increasing very rapidly particularly increasing rate of rapes and murders
===== Loaded model: classy.dat =====
===== Instance created with reference dataset =====
@relation 'Test relation'

@attribute text string
@attribute class {positive,negative}

@data
'I agree with Capital punishment today we find criminality is increasing very rapidly particularly increasing rate of rapes and murders
===== Classified instance =====
Class predicted: positive
```

The model correctly classified this test instance as positive.

```
Problems @ Javadoc Declaration Console
<terminated> text2 [Java Application] C:\Program Files\Java\jdk1.8.0_05\bin\javaw.exe (May 13, 2015, 3:53:16 PM)
===== Loaded text data: C:\Users\Vadehra\Desktop\smstest.txt =====
I think capital or death punishment should be put to an end this is a topic going in global world and most of the countries are in favor o
===== Loaded model: classy.dat =====
===== Instance created with reference dataset =====
@relation 'Test relation'

@attribute text string
@attribute class {positive,negative}

@data
' I think capital or death punishment should be put to an end this is a topic going in global world and most of the countries are in favor
===== Classified instance =====
Class predicted: negative
```

The model correctly classified this test instance as negative

```
Problems @ Javadoc Declaration Console
<terminated> text2 [Java Application] C:\Program Files\Java\jdk1.8.0_05\bin\javaw.exe (May 13, 2015, 3:55:24 PM)
===== Loaded text data: C:\Users\Vadehra\Desktop\smstest.txt =====
As rightly said An eye for eye makes the whole world blind Those who commit rape or heinous crimes like that need more suffering than death My point of
===== Loaded model: classy.dat =====
===== Instance created with reference dataset =====
@relation 'Test relation'

@attribute text string
@attribute class {positive,negative}

@data
' As rightly said An eye for eye makes the whole world blind Those who commit rape or heinous crimes like that need more suffering than death My point of
===== Classified instance =====
Class predicted: negative
```

The model correctly classified this test instance as negative.

CHAPTER 9:

CONCLUSION

Emotion Detection may be seen as an essential field of research in human-pc interplay. A enough quantity of work has been done via researchers to come across emotion from facial and audio statistics whereas recognizing emotions from textual records is still a fresh and warm research area.

The record demonstrates that it's far viable to extract information approximately someone's sentiments from the textual content they write on social networking sites with high accuracy. The approach applied helps to get statistics about the users' sentiment polarity (high quality, impartial or poor) in step with the messages they write.

The class technique implemented follows a lexical-based totally method.

Adaptive and advice structures in preferred can take benefit of knowing the customers' sentiments at a sure time, in addition to sizeable emotional adjustments with appreciate to their normal state. However, asking the customers approximately their sentiments is intrusive, may be bothering and, furthermore, in a few contexts there may be a high chance that they do not admit poor sentiments, because of, e.G., their cultural historical past or the need of social approval. In the academic context, specially, it isn't always fairly probably that a scholar immediately transmits the teacher his/her feelings towards a topic or methodology whilst they're negative. The technique to get this statistics is typically more difficult (e.G., anonymous/ distinctive surveys at the stop of the semester, or teachers receiving feedback through class delegates in the course of the direction, at most). The work presented in this file demonstrates that it's miles possible to extract it from the messages the scholars write on social networking websites.

This can be considered as enter for adaptive e-getting to know systems to provide sentiment-based totally adaptation, making it feasible, e.G., to suggest the most appropriate sports to be executed through every student at a sure time in line with his/her sentiment at that point; another software of this paintings in the learning context deals with extracting remarks for instructors about the emotions of their students toward their guides or teaching methodologies

CHAPTER 10:

FUTURE WORK

- One potential improvement to the project will be the ability to detect not just the user's emotions but also the significant emotional changes over a period of time, like a day or a week. This can be useful in adaptive E-learning systems to assess what effects the subjects taught are having on the student's state of mind.
- Another improvement is the ability to rectify and detect textese.
- One other change could be the ability to detect the emotion behind a word based on the context it has been used in.
- The emotion of a word also changes based on its position in the sentence. A potential improvement would be to take this position under account.

CHAPTER 11:

REFERENCES

- [1] Walaa Medhat, Ahmed Hassan and HodaKorashy, “**Sentiment analysis algorithms and applications: A survey**”, Ain Shams Engineering Journal, April 2014
- [2] Alvaro Ortigosa, José M. Martín and Rosa M. Carro, “**Sentiment analysis in Facebook and its application to e-learning**”, Computers in Human Behavior, 2013
- [3] XIONG XiaoBing, ZHOU Gang, HUANG YongZhong, CHENHaiYong and XU Ke, “**Dynamic evolution of collective emotions in social networks**”, July 2013, Vol. 56
- [4] Shiv Naresh Shivhare and Prof. Saritha Khethawat, “**Emotion Detection from Text**”, Department of CSE and IT, Maulana Azad National Institute of Technology, 2011, Bhopal, Madhya Pradesh, India
- [5] Cecilia OvesdotterAlm, Dan Roth and Richard Spooth, “**Emotions from text: machine learning for text-based emotion prediction**”, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), October 2005, pages 579–586
- [6] Bo Pang, Lillian Lee and ShivakumarVaithyanathan, “**Thumbs up? Sentiment Classification using Machine Learning Techniques**”, Proceedings of EMNLP 2002, pp. 79–86
- [7] Douglas R. Rice and Christopher Zorn, “**Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies**”, Prepared for presentation at the New Directions in Analyzing Text as DataWorkshop, 2013, Version 0.1, September 2013
- [8] Andrew McCallum and Kamal Nigam, “**A Comparison of Event Models for Naive Bayes Text Classification**”

Plag_Report

ORIGINALITY REPORT

3%

SIMILARITY INDEX

2%

INTERNET SOURCES

2%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1

www.coursehero.com

Internet Source

1%

2

Walaa Medhat, Ahmed Hassan, Hoda Korashy. "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, 2014

Publication

1%

3

G Krishna Chaitanya, Dinesh Reddy Meka, Vakalapudi Surya Vamsi, M V S Ravi Karthik. "A Survey on Twitter Sentimental Analysis with Machine Learning Techniques", International Journal of Engineering & Technology, 2018

Publication

<1%

4

Submitted to Jaypee University of Information Technology

Student Paper

<1%

5

Lecture Notes in Computer Science, 2004.

Publication

<1%

Exclude quotes On

Exclude matches < 15 words

Exclude bibliography On

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 17 June 2021

Type of Document (Tick): ~~PhD Thesis~~ ~~M.Tech Dissertation/ Report~~ B.Tech Project Report Paper

Name: AAYUSH VERMA Department: CSE Enrolment No 171259

Contact No. 9419226518 E-mail. aayushv.99@gmail.com

Name of the Supervisor: Dr. YUGAL KUMAR

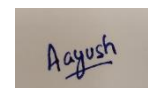
Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): Personal Care Monitoring System

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/ revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages = 58
- Total No. of Preliminary pages = 7
- Total No. of pages accommodate bibliography/references = 1



(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at **3 (%)**. Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor)

Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	<ul style="list-style-type: none">• All Preliminary Pages• Bibliography/Images/Quotes• 14 Words String		Word Counts	
Report Generated on		Submission ID	Character Counts	
			Total Pages Scanned	
			File Size	

Checked by
Name & Signature

Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com