ONLINE REVIEWS SPAM DETECTION USING FEATURE SELECTION

Enrolment No. (s) - **9913103629**
9913103666
Name of Student (s) - **KawaldeepKaur**
Prachi Gupta
Name of Supervisor - **Mr. Himanshu Mittal**

May – 2017

Submitted in fulfillment ofthe Degree of
Bachelor of Technology

in

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY.

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

TABLE OF CONTENTS

DECLARATION

We hereby declare that the submission of this report contains our own work to the best of our knowledge and it contains no matter previously used or published by any other person nor it has been accepted for anyaward for anyotherdegreeor diploma of university or any institute of higher learning except the places where due acknowlegement has been made in the text.

Place : JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

Date :

Signatue:

Name : Kawaldeepkaur (9913103629)

Prachigupta (9913103666)

CERTIFICATE

This is to certify that the project titled "**Online Reviews Spam Detection Using Feature Selection**" submitted by "**Prachi Gupta and KawaldeepKaur"** in fulfillment for the degree of bachelor in technology of jaypee institute of information technology , Noida ,has been throughout my supervision. This work is there own and is not being submitted for award by any other institute or university of this or any other degree or diploma.

Signature of Supervisor ……

Name of Supervisor Mr. Himanshu Mittal

Designation Assistant Professor(CSE)

Date

ACKNOWLEDGEMENT

We would like to place on record our deep sense of gratitude to Mr Himanshu Mittal, our major project mentor ,for his generous guidance, help ,useful suggestions, stimulating guidance, continuous encouragement and supervision throughout the course of present work.

We also wish to extend our thanks to other classmates for their insightful comments and constructive suggestions to improve the quality of this project work.

KawaldeepKaur (9913103629)
PrachiGupta(9913103666)
Date ……..

SUMMARY

The speedy growth in the number of ecommerce websites ,has made the web ,an intensive source of reviews on anything. Since there's no scrutiny relating to the standard of review written ,anyone can easily write something that is false and hence is made to mislead the users which are called review spams. There has been a rise within the variety of deceptive review spams. Fake reviews that appear real however are deliberately made up to have an effect on the user. We have used supervised methodologies during our work for spotting the review spams. During our research we found that various improved techniques have been proposed till date for finding most effective feature set for carrying out the analysis for best accuracy. Sentiment analysis is being conducted and its results have conjointly been used and integrated into the spam review detection . some acknowledged classifiers have been used on the labeled dataset for best performance results. From our results , compute the best feature set for getting the best accuracy using minimum number of features rather than all the features which helps in reducing time. Therefore for feature selection ,modified cuckoo search algorithm is being used. Further SVMmachine learning algorithm is implemented on our labeled dataset using the selected features only. Hence to lead to detection of spam and spammers in the product review set. Then there is comparison between the results found on accuracy and time before without using cuckoo search algorithm and with cuckoo search algorithm. It proposes a Cuckoo search approach that is applied to the featuresobtained with best accuracy. The dataset on hotel reviews is being collected from Yelp , having reviewer data is used for analysis. The proposed algorithm analyses the dataset and makes several useful observations.

---------- ----------

Signature of Students Signature of Supervisor

KawaldeepKaur (9913103629) Name Mr.Himanshu Mittal

PrachiGupta(9913103666)

Date Date

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 : INTRODUCTION

1.1 GENERAL INTRODUCTION

We as a whole know spam (undesirable email promoting) is a hassling matter. In the event that you are an email client, it squanders your time. In the event that you are a site proprietor or a server administrator, it takes valuable assets from your email server. Those whose work is identified with the Internet realize that battling spam is a mouse-and-feline race between its senders and email suppliers, however we don't typically know how does this spam/hostile to spam war works or occur. I thought a short clarification can deal with it. Online audits are frequently the essential calculate a client's choice to buy an item or benefit, and are a profitable wellspring of data that can be utilized to decide popular feeling on these items or administrations. As a result of their effect, makers and retailers are profoundly worried with client input and audits. Dependence on online audits offers ascend to the potential worry that Cretans may make false surveys to misleadingly advance or debase items and administrations. This practice is known as Opinion (Review) Spam, where spammers control and toxin surveys (i.e., making fake, untruthful, or misleading audits) for benefit or pick up. Since not every online audit are honest and dependable, it is vital to create procedures for recognizing survey spam. By removing important components from the arrangement of surveys, it is conceivable to lead audit spam location utilizing different machine learning procedures. Also, commentator data, aside from the content itself, can be utilized to help in this procedure. As the Internet keeps on developing in both size and significance, the amount and effect of online audits persistently increments. Audits can impact individuals over an expansive range of businesses, yet are especially imperative in the domain of web based business, where remarks and surveys with respect to items and administrations are regularly the most helpful, if by all account not the only, route for a purchaser to settle on a choice on regardless of whether to get them. In this way there is an incredible need to grow such works which can vary genuine and fake audits.

1.2 PROBLEM STATEMENT

Our main goal is to devise automated methods to detect review spams in online reviews using feature selection from set of multiple reviews. We obtain the most apt datasets for the study of this problem. We try to obtain the feature sets that can very well represent and differentiate

the spams reviews and the non spam reviews. We implement it using nature inspired algorithms and further applying feature selection on it. We also deduce the result using support vector machine. We apply support vector machine to calculate the best score or measure with different features .We also use Cuckoo Search algorithm to obtain the best population in our review spam detection. Lastly, we get the most optimum result of best features.

1.3 EMPIRICAL STUDY

FIELD SURVEY

Web crawlers turned into a true place to begin data procurement on the Web. In spite of the fact that because of web spam marvel, query items are not generally on a par with craved. In addition, spam develops that makes the issue of giving superb hunt considerably additionally difficult. Current important issues other than online surveys ,where spam can be recognized are as per the following:

Email Spam : Direct mail messages are utilized to target singular clients in Email Spam. The rundown for email spams is frequently arranged by filtering the web for Usenet postings, web hunt of locations and in addition taking of web locations.

Remark Spam : Another classification incorporates, remark spam which is generally utilized by spammer by posting remarks for their accursed reason.

Texting Spam : This kind of spam makes utilization of texting frameworks. Texting is a for of visit based direct correspondence between two individuals continuously, utilizing either PCs or whatever other gadgets. The system imparts messages just as content. It is extremely basic on numerous texting frameworks, for example, Skype.

Garbage Fax : Junk fax is a methods for showcasing by means of spontaneous promotions that are sent through fax. So the garbage faxes are essentially what might as well be called a spam mail. It is a medium of telemarketing and promotions.

Spontaneous Text Messages Spam or SMS Spam : This kind of spam (SMS) is difficult to channel. Because of the minimal effort of web and quick advance as far as innovation, it is presently effectively conceivable to send SMS spams at fundamental sums utilizing the Internet's SMS entries. It is quick turning into a major test that should be overcome.

Person to person communication Spam : Social Networking spam is focused for the standard clients of the long range interpersonal communication sites, for example, LinkedIn,

Facebook, Google+ or MySpace. It regularly happens that these clients of the informal communication web administrations send coordinate messages or web connects that contain inserted joins or malevolent and spam URLs to different areas on the web or to each other. This is the manner by which a social spammer assumes his part.

TOOLS

• **MATLAB:** is a numerical computing environment and fourth generation programming language.

• **PYTHON:** Python is a widely used high level dynamic programming language.

HARDWARE & SOFTWARE DEPENDENCIES

• **Operating System:** Windows 9x/XP or higher ,Windows ME

• **Processor:** Pentium 3.0 GHz or higher

• **RAM:** 250 Mb plus

• **Hard Drive:** 10 GB plus

• MATLAB

• IPython Console Natural Language Toolkit

1.4 APPROACH TO PROBLEM IN TERM OF TECHNOLOGY AND PLATFORM TO BE USED

NATURE INSPIRED ALGORITHM

We are using nature inspired algorithm , to implement the solution to spam detection .further we will be comparing our result with the support vector machine results , to check the accuracy of our solution.

Nature-inspired algorithms are considered among the most intense calculations for improvement. These incorporate numerous calculations, for instance ,the PSO calculation ,it scans for the space of the target elements of individual specialists by changing the directions , called particles, as the piecewise ways shaped in a semi stochastic way by positional vectors. Then there is Firefly Algorithm which can be modified to take care of multi target enhancement issues. Likewise , the use of the firefly algorithms in mix with alternate calculations may frame an extremely energizing zone for further inquires about.

The Evolutionary algorithms optimizing agents are worldwide enhancement techniques and these calculations scale well to higher dimensional issues. These calculations are powerful concerning the loud assessment capacities, and furthermore the treatment of assessment capacities which don't yield a sensible outcomes in given timeframe is clear. In this venture fundamentally we will work with Cuckoo Search Algorithm. The calculation proposition was roused by the commit brood parasitism of a portion of the cuckoo species by laying their eggs in the homes of other birds(of have species). Some host feathered creatures can connect with direct clash with the barging in cuckoos.

CUCKOO SEARCH

Cuckoo search was created in 2009 by xin-she Yang and suash Deb. The algorithm proposition was propelled by the commit brood parasitism of a portion of the cuckoo species by laying their eggs in the homes of other birds(of have species). Some host cuckoos can draw in direct clash with the encroaching cuckoos. For instance, if a host fowl finds that the eggs are not their own, then it will either discard these outsider eggs or basically relinquish its home and construct another home elsewhere. Foe example ,Some cuckoo species like the New World brood-parasitic Tapera have developed such that female parasitic cuckoos are frequently very specialized in the mimicry in hues and pattern of the eggs of chosen host species .Cuckoo search glorified such reproducing conduct, and in this manner can be connected for different optimization problems.

SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) classifier is particularly represented by a separating hyperplane. Suppose, we are given labeled training dataset, the algorithm uses supervised learning method thus producing a hyperplane that is the most optimized. This optimized hyperplane then classifies dataset from test set.

Thus, we need to figure out a straight line that separates 2D points in a linear fashion which are distributed among the two classes.

In the process of finding an optimal straight line, if it ends up being close to any point, it will be a bad generalization and might be sensitive to noise and thus incorrect. Thus, our objective will be to be able to get a straight line that is farthest possible from the class points while dividing the class. The goal of our SVM classifier is to find a hyperplane giving farthest minimum distance

between the training class points. We also find something called ”margin” in the SVM classifier theory that is twice this separating distance. Finally this hyperplane that we have found, tends to maximize out training data’s ”margin”.

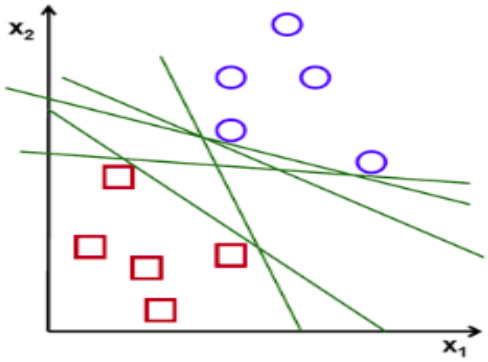

Fig 1.1. Linearly separable set of 2-D points

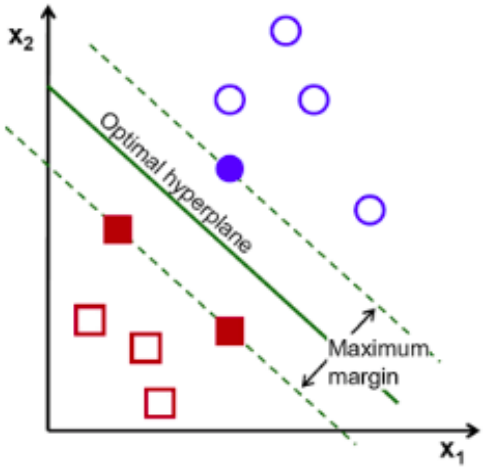

Fig.1.2. Optimal hyperplane using svm classified

1.5 TABULAR COMPARISON OF THE PROBLEM FRAMED AND OTHER EXISTING APPROACHES.

Table1.1. Comparison of Review spam detection techniques with labeled data

CONCEPT	FEATURES	LEARNER	RESULT
duplication of Text	reviews ,reviewers and product centric	Logistic_ regression	78.0%-accuracy
Similar content	LIWC and bigram	Support vector machine	89.3%-accuracy
Similar Text	Text Review	Support vector machine	81.0%-precision
Similar products features	Features of product	cos similarity	42.6%-precision
Stylometrics	lexical and syntactical	Support vector machine	84.2%-F_score
Content_similarity	behavioral +bigrams	Support vector machine	85.1%-accuracy
Ontology	Ontological features	conditional filtering	76%-precision
content review(taking negative reviews)	n-gram	Support vector machine	85%-accuracy
content similarity and sentiment polarity	linguistic features and Unigrams	decision_tree	92.12%-accuracy

CHAPTER 2: LITERATURE SURVEY

Various analyses and research are made in the of field of Spam detection. Many researchers have proposed Support Vector Machine, Naive Bayes, Multilayer Neural Networks, spam filtering methods using different classifiers including hybrid models, observational assessment of the viability of techniques for highlight choice for lessening the list of capabilities size in audit spam location and some more. Directed learning is the most usually utilized procedure in assessment spam identification. In spite of the fact that conclusion spam (or fake audit) discovery has pulled in significant look into consideration as of late, the issue is a long way from illuminated. One key reason is that there is no huge scale ground truth marked dataset accessible for model building. As the quantity of survey destinations and audits develops, strategies for reducing the quantity of components is getting to be noticeably essential, since the list of capabilities sizes can develop past what can be taken care of by customary machine learning procedures. Since when the spammers compose the fake surveys, they have a tendency to portray an item utilizing some extraordinary element words and wistful words. It is useful for the fake audit location.

2.1 SUMMARY OF RELEVANT PAPERS

Research Paper 1

1.Title: A Binary Particle Swarm Optimization Algorithm and Its Applications on Feature Selection Problems and Knapsack

2. Authors: Bach H. Nguyen, Bing Xuie and Peter Andreaae

3.Year of Publication :2012

4. Publishing Details: It is published in 20th Asia Pacific Symposium, Canberra, Australia, November 2016.

5.Summary: In the paper, we found that it introduced new concept to the nature inspired algorithm i.e. BPSO on the basis of which a new updating mechanism introduced a new concept to BPSO based on which they developed a new updating mechanism called SBPSO. They compared this newly developed BPSO by two well known binary problems i.e. feature selection and knapsack problem. Through the comparison analysis it was found that SBPSO has better performance than the knapsack problem on all datasets while in feature selection, due to selected

number of features the accuracy is degraded . Hence SBPSO provided best results. Also it detected that when there are large number of features , then the introduced momemtum by the training accuracies and evolutionary processes helped SBPSO to improve over PBPSO. While SBPSO is a costly algorithm over the PBPSO due to its complex computation and also as it needs to maintain the current Lifevector .

The work was mainly focused on introducing the momentum concept to BPSO ,more works are need to be conducted to optimize its parameters. Also ,the proposed algorithm is only for binary problems .the author wished to extend this work for SBPSO for generic problems and also to multi objective BPSO algorithm to consider trade off and optimize multiple conflicting objectives.

6. WebLink: http://link.springer.com/chapter//10.11007/978-3-319-49049-6_23

Research Paper 2

1. Title : An Efficient Hybrid Algorithm using Differential Evolution for Data Clustering and Cuckoo Search.

2. Authors: AsgaraliBouyer, HabibGhafarzadeh and OmidTarkhaneh

3. Year of Publication: 2015

4. Publishing details

5. Summary: In this paper we studied about an effective hybrid algorithm for data clustering proposed by the author. The proposed algorithm modifies the cuckoo search algorithm which benefits from DE to produce new nests in CS and it also benefits from Mantegna Levy flights to boost the local search and for better performance results. The algorithm used in the paper seems to outperform the other algorithms in terms of better total within cluster variance than the PSO, CS, GSA ,DE, BH and BB-BC. It also tested HCSDE with other six well known real datasets from USI machine learning repository. It was clear from the results shown in the tables that the proposed algorithm has better results for accuracy and error rate than the other algorithms. However , the proposed algorithm may get in into local minima in few cases . So , improving the features of the algorithm is considered in there future work.

6.Weblink: http://www.indjst.org/index.php/indjst/article/view/601466

Research Paper 3

1.Title: Review spam detection using machine learning technique.

2. Authors: M.Crawford, T.M. Khosh goftaar, Joseph D. Prusaa, Aaran N. Rechter and H.A. Najada

3.Year of Publication: 2015

4. Publishing details: Published Online :05 October 2015,Journal of Big Data – a SpingerOpenJournel© 2015 Crawford et al.

5. Summary : This paper incorporates The Feature Engineering for Review Spam Detection. The section provides a brief on feature engineering, for both reviewer centric and review centric spam detection. It provides proper discussion on the topics . In the review centric review spam detection, it analyzes current research using all the machine learning techniques i.e supervised ,unsupervised and semi supervised. And for the reviewer centric review spam detection ,an overview is provided on the section using reviewer centric features. It contains discussion and comparison in comparative anakysis and suggestions for different methods proposed.

6. Web link: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0029-99

Research Paper 4

1.Title of paper: Sentiment analysis using review data of product.

2.Authors: XingFang andJustinZhan

3.Year of Publication: 2015

4.Publishing details : Journal of Big Data – a SpingerOpenJournel

5.Summary : This paper basically worked on handling sentiment analysis, i.e sentiment polarity classification. The author proposed an algorithm for negative sentiment phrase recognition in the

dataset. It contained the proposed mathematical approach for the sentiment score calculation. In one section of the paper a feature vector generation method is shown .While in the second phase two experiments on sentiment polarity are being performed. In last section, all the performances for the three classification models are evaluated. Then the comparison is made on the basis of the results obtained from them. The dataset used is collected from amazon on product reviews.
In the paper , the author inspects all the product reviews before posting them and every review is checked for its ratings that should be on the basis of truth. This is done to overcome the flaws in the product reviews.

6.Web link: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

Research Paper 5

1.Title of paper: Comparative opinion mining: A review

2.Authors: KasturiDewiVarathan, Anasthasia Giachanu and F. Crestanii

3.Year of Publication: 2016

4.Publishing details: Published online: 3 August 2016,by- Swiss Secretariat of Research,Innovation and Education.

5.Summary : This article concentrates on relative conclusion mining in which the principal segment of the article gives a short presentation on sentiment mining and near assessment mining. The fundamental substance of this article concentrates on similar supposition mining strategies and mining near components, which incorporate mining similar sentences, element discovery, connection recognition, and highlight identification. Assets accessible for relative conclusion mining are additionally talked about to help future scientists utilize these assets. Concentrates put forth on similar expressions all in all content, for example, web reports, news, or whatever other logical content are not inside the extent of this review since they may contain truths as opposed to sentiments. At last, it display the distinctive execution measures that are as often as possible utilized as a part of the writing of near assessment mining

6.Web link: http://onlinelibrary.wiley.com/doi/10.1002/asi.23716/full/

Research Paper 6

1.Title of paper: Spotting Fake Reviews by Collective Positive Unlabeled Data Learning

2.Authors: Huay- Li, Bings Liu, Xiaokai W, Zhiyuan Chin and Jidang Shaow

3.Year of Publication: 2014

4.Publishing details : Data Mining ,IEEE International Conference in 2014 , Shenzhen ,China

5.Summary : It The paper proposes a aggregate classification algorithm MHCC (multi typed heterogeneous collective classification)to recognize fake reviews. In the defined heterogeneous network on reviews , users and IP addresses. Also , exploitation of the relational features is restricted by the small size of training data for classification. The MHCC algorithm is extended to the collective PU learning model(CPU) to adapt to the positive and unlabeled data. The unlabeled data is treated as negative only at stage of initialization. But it later runs repeatedly to sort both negative and positive reviews from unlabeled dataset. As once the classifier is learned then it starts assessing its classification results so as to give proper positive and negative based instances on which further classifiers are trained .

6.Web link: http://ieeexplore.ieee.org/document/7023420/

Research paper 7

1. Title of the paper : Detecting Product Review Spammers using rating Behaviours

2. Author : Ee-PengLim, Bing-Lin, Nitina Jindal, Hady-W.Lauw, , Viet-An-Ngygen

3. Year of publication: 2010

4.Publishing details : Publication of this paper was done on 19th ACM international conference which was for knowledge management and information management at Toronto in Canada .

5. Summary : In the paper ,it is basically detecting the reviews by users who generate spam

reviews. It identifies many characteristics of spammers and use this behavior to detect to review spammers. The characteristic behaviours like spammers may try to deviate from the other reviewers and misguide the users by giving wrong ratings of the product so as to promote or demote it. Next ,spammers to maximize their impact may target product groups or special groups. It proposed methods for measuring the degree of spams for the reviews and apply that on dataset. Afterwards , the highly suspicious reviews are being found out using the web based spammer evaluation software. The proposed solution comes out to be correct in finding the spammers and also outperform other baseline methods based on helpful cots only.i in the paper it was further detected that detected spams have more significant effect on results than the unhelpful reviews.

6. Weblink : http://www.cs.uic.edu/liub/publications/cikm-2010-final-spam.pdf-.

2.2 RESEARCH PAPERS INTIGRATED SUMMARY

As we know now a days not every online survey are reliable and honest, so it is imperative to create procedures for distinguishing spam in audits. We can do that by separating important elements from the utilizing Natural Language Processing ,it is conceivable to direct spam discovery on surveys utilizing different machine learning systems. Moreover, analyst data can be utilized to help in this procedure , separated from the itself. We have done the overview on the unmistakable machine learning strategies that have been proposed to take care of the issue of survey spam identification and furthermore the execution of various methodologies for location and characterization of audit spam. The dominant part of flow research is being centered around directed learning techniques, which require marked information, a shortage with regards to online survey spam.

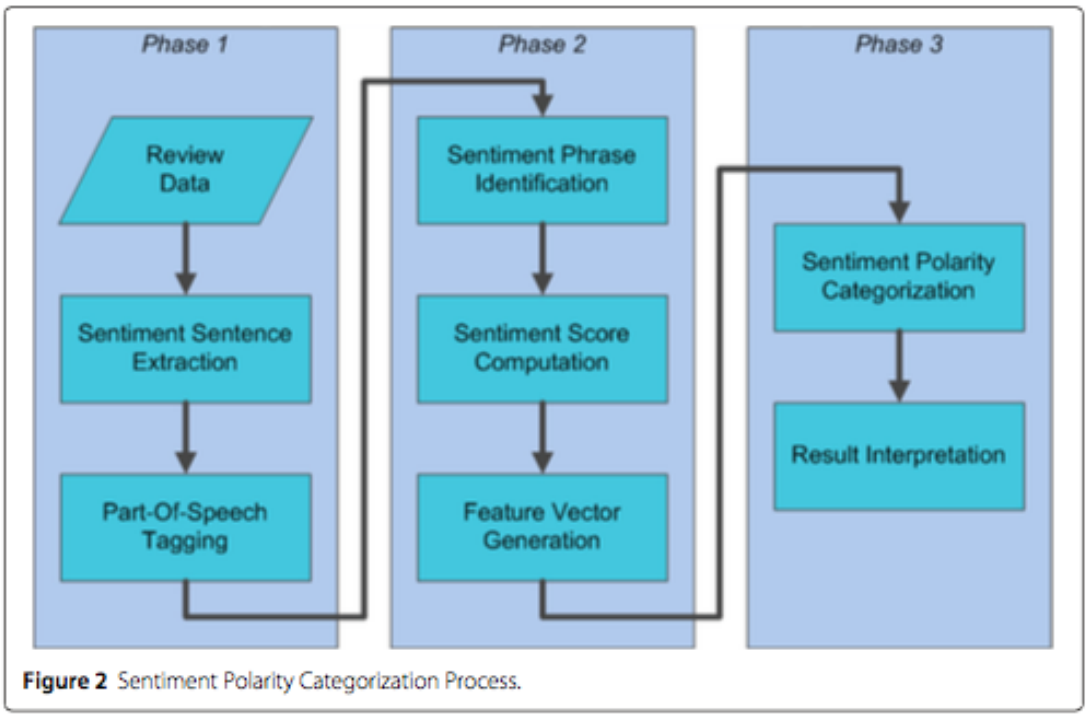

Fig 2.1. Diagramatic representation of phases

The flowchart that the proposed procedure for arrangement and the blueprint of the exploration paper. The commitments for the most part fall into Phase 2 and 3.

In Phase 2:

1) A calculation is proposed and actualized for negative expressions recognizable proof;

2) A numerical approach is proposed for notions score calculation;

3) A component vector era strategy is exhibited for feeling extremity arrangement.

In Phase 3:

1) Two supposition extremity classification tests are separately performed in view of sentence level and audit level;

2) Performance of three arrangement models are assessed and thought about in view of their exploratory outcomes.

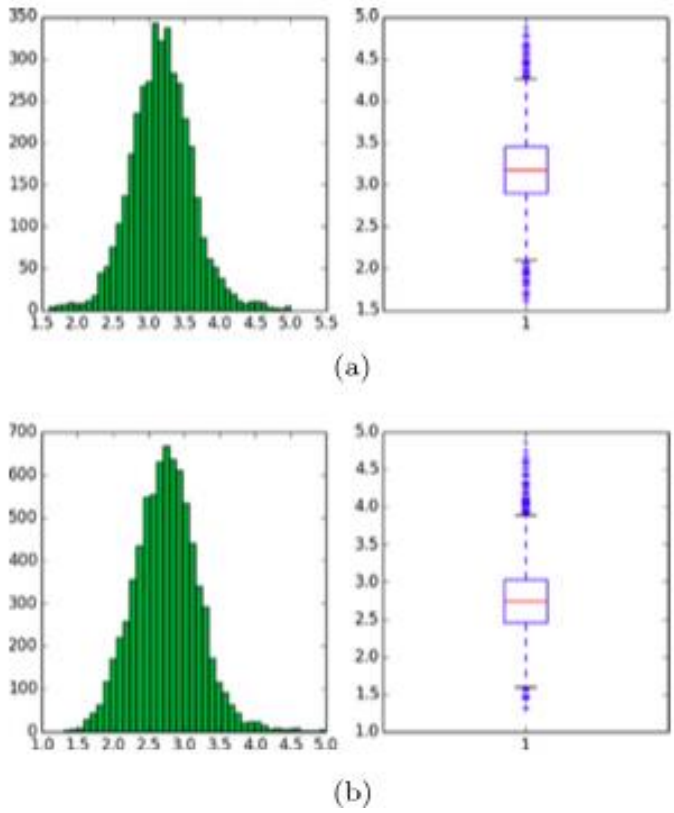

Fig 2.2. Sentiment_score (a) positive words (b) negative words

Subsequently, the notion score data for positive word tokens is appearing in Figure. The histogram outline depicts the circulation of scores while the case plot graph demonstrates that the middle is over 3. So also, the case plot outline in Figure 2.2(b) demonstrates that the middle of opinion scores for negative word tokens is lower than 3. Truth be told, both the mean and the middle of positive word tokens do surpass 3, and both qualities are lower than 3, for negative word tokens (Table).

Table 2.1.Analytical information

TOKEN	TYPE	MEAN	MEDIAN
Positive_word	Token	3.17	3.15
Negative_word	Token	2.86	2.69

In our exploration papers we have concentrated different courses for spam identification ,for certifiable data , different systems and calculations used to for doing likewise ,each examination paper had its own particular new proposed technique to take care of the issue.

For identifying spams , we need to do different calculations in light of supposition investigation, common dialect preparing on different online audits. Web mining ,machine learning , enormous information and arrangements frame the nuts and bolts of examination to be done.

Overviews were directed and results were discovered utilizing calculations like grouping , gullible Bayes ,bolster vector machine and being analyzed . Estimation investigation or assessment mining is one of the real undertakings .One of the proposed arrangement in the exploration papers on conclusion mining was on the accompanying examination

The fake negative analysts are seen to over-create terms delineating negative feelings (e.g., horrendous, baffled, and so forth.) when contrasted with the honest audits. Also, invented positive commentators over-delivered terms delineating feelings of energy (e.g., wonderful, exquisite, and so forth.). In this manner, fake inn analysts overstate the notion.

Tricky or fake audits have been found to contain an awesome rate of words demonstrating positive opinions than the genuine positive surveys. Also ,beguiling negative audits contain more negative terms than the honest to goodness negative surveys

For proposed technique, they have utilized opinion score as an element as a result of the variety between assumption extremity and rating of surveys.

Steps

i. Extract highlights/perspective things from each sentence in the audit.

ii. We locate the relating opinion words show in the sentence.

iii. Strength of the feeling word on the component diminishes with the separation from the element word.

The investigation of audit spam location in light of various element models, e.g. etymological, POS, n-gram, and slant. In pragmatic, these elements in mix beat the adequacy of each of them in particular. Subsequently a combinatorial investigation is likewise detailed.

At long last, investigation of our proposed bound together components model is exhibited.Analysis Linguistic features

This Linguistic elements examination works averagely. In spite of the fact that, it has been have watched that this investigation is equivalent to the grouping done physically by human annotators in ordering the same dataset. We have found that people have an exactness level of under 60 % for the same dataset grouping undertaking. While, When different gatherings were made a request to group the dataset, their simultaneousness of results was entirely low. Along these lines, semantic elements model is tuned in to the human instinct in misleading surveys discovery.

Analysis of POS features

POS Features investigation likewise gives us a normal outcome appeared yet its not in the same class as the outcomes given by the etymological examination. Consequently in the following segments, we consolidate the two techniques.

Analysis of N_gram features

The outcomes acquired from n-gram content characterization is appeared in Fig. 2.3.The precision levels gotten in n gram include investigation is genuinely superior to anything the ones acquired from phonetic and POS models. Taking after perceptions can be made about the same:

i.it is watched that spammers utilize an arrangement of words every now and again in contrast with the honest to goodness audit journalists. This property is sufficiently useful for us to characterize spam conduct.

ii. We find that spammers and non-spammers may have utilized comparative words, yet the recurrence of its utilization from the word-sets has a tremendous effect.

iii. We can utilize the n-gram display as a rule in a wide range of situations, not to mention lodging surveys as the fundamental thought stays same and this technique functions admirably on a wide range of datasets .

iv. The assessment scores unquestionably realize an expansion in the exactness got when joined with alternate components.

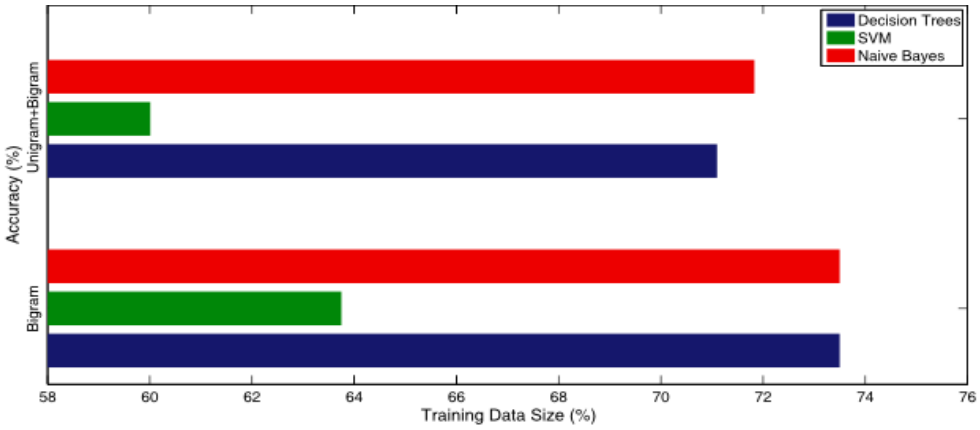

Fig 2.3. N gram analysis

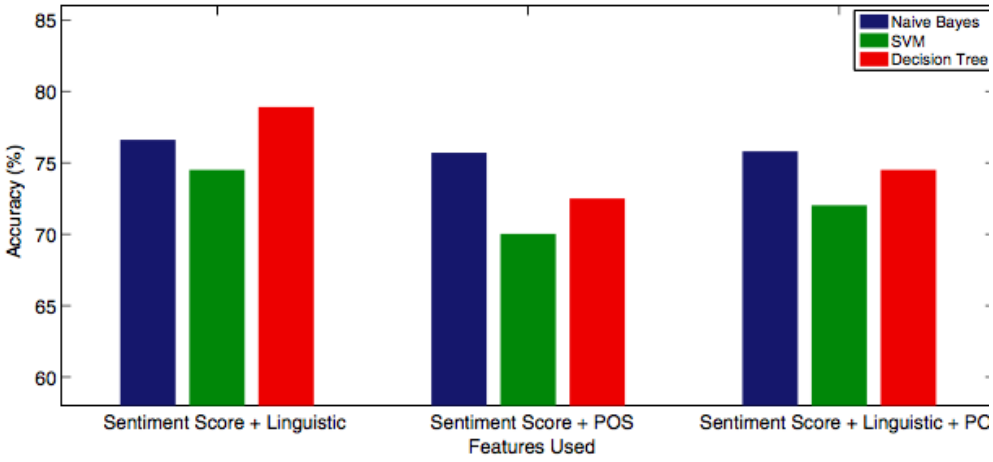

Fig 2.4. Sentiment feature analysis

Table 2.2. Unified feature models analysis

Features	Classifier	Precision
Sentiment_score + linguistic_feature +unigram_feature	Naïve_Bayes	90
	SVM	83.25
	Decsion tree	91.22

CHAPTER 3: ANALYSIS , DESIGN AND MODELING

3.1 OVERALL PROJECT DESCRIPTION

3.1.1 COLLECTION OF DATASET

Yelp: Yelp.com is a popular review website which is crowd-sourced and reviews local stores and brands. Here, users can also interact with one another just like in social networking sites. It is more popular in metropolitan areas as a review site. Here, users can rate products or services such as restaurants, mobiles, etc. Star ratings between 1 to 5 can be given by users. After that they could descriptively write about a product or service. Also, users could check-in, just like Tripadvisor.com, into a restaurant, hotel or a location that they are visiting. Yelp gets about 132 million visitors on a monthly basis and about a total of half a billion reviews. Although, Yelp does not give away datasets to the public, we can scrape user information and reviews from their website. though Yelp does not provide its dataset publicly, the reviews and user information can be scraped from the site itself. Bots and scripts can be used to scrape the data as they are allowed with low security so as to get more penetration in search engine results.

3.1.2 DATASET DESCRIPTION

We assembled a gathering of a sum of 4400 surveys from the sources said above.
These audits were for 20 Chicago-based lodgings. The accompanying are the elements of each audit:

1. An exceptional ID for each survey for audit identification
2. The inn name about which survey has been composed
3. The substance of the audit
4. The extremity of the audit, as in whether it depicts positive or a negative slant
5.The paired mark for delineating whether the audit is a spam or not

3.1.3 FEATURE COLLECTION

Linguistic Characteristics as Features

Table 3.1.linguistic characteristics as features

Feature ID	Description of features	Linguistic features
F.A	Number of word used.	Quantity as a feature
F.B	Average number of words per sentence used.	Complexity as a feature
F.C	Number of diverse/different words present.	Diversity as a feature
F.D	Brand names present.	Brand name
F.E	Percentage of number of characters and number of words being used.	Average length of words used
F.F	Number of digits present.	Digits as a feature

Class Identification: POS Tagging- a Feature

We look at the relationshipexisting amongst honest to goodness and misleading audits in our way to deal with find beguiling survey spams. In view of the POS labels we figure 9 include values for each survey these are specifically, thing, pronoun, modifier, intensifier, verb, determiner, prepositions,coordinating conjunctionsand predeterminers.

To look at exhibitions of classificationmodels created and other robotized algorithmic procedures these POS label characteristics give a gauge.

Table 3.2. POS tagging as feature

Features	**POS tagging.**	**Description of reviews**
F.G	**AD**	**Count of Adjectives used**
F.H	**NU**	**Count of Nouns used**
F.I	**PRE**	**Count of Preposition used**
F.J	**DTE**	**Count of Determiners used**
F.K	**VBS**	**Count of Verbs used**
F.L	**ADV**	**Count of Adverbs used**
F.M	**PN**	**Count of Pronouns used**
F.N	**CW**	**Count of Connector Words used**
F.O	**FPN**	**Count of First Pronouns used**

Text Categorisation: N-gram as a Feature

Steps:

1. For fusing N-grams as a component, we consider unigram and bigram as ourfeature sets, with the N-grams in lower case and unstemmed.

2. We keep up a lexicon for our unigrams highlights acquired fromthe preparing dataset.

3. Presently, from the test site, each audit taken is then part into the correspondingN-grams. For every N-gram, its relating score is checked.

4. The score depends on either nearness in a spam/non-spam set, or its absence,taken in 0s in the particular cases.

5. At long last, we compute the aggregate scores to get a thought whether the test audit ismore comparable with spam set or the non-spam set to have the capacity to figure out whether it is honest to goodness or fake.

Presently, this score is utilized to display our classification dataset.

Table 3.3. N-gram as a Feature

Feature Number	Description of feature	n-gram feature
F.P	Score indicates how much of the words of a review entered are same as those to the spams reviews	Spam_Hit_Score
F.Q	Score indicates how much of the words of a review entered are same as those to the spams reviews	Non_Spam_Hit_Score

3.1.3 REQUIREMENT SPECIFICATION

Via supervised learning spam filtering has traditionally relied on extracting spam signatures i.e., using emails explicitly labeled as spam or harm. Such supervised learning is labor-required and cost not effective also most important thing is that it cannot adapt to new spamming behavior quickly. So, to filter that can more effectively identify new spamming behavior we study the feasibility of unsupervised learning-based spam .

Features Observed: Linguistic n-grams have been shown to be useful for spam detection. We tend to construct the behavioral features from various abnormal behavioral patterns of reviews and reviewers.

The Following features have been observed:

Content Similarity: Spammers normally post fake encounters. They frequently post surveys which are copy or close copy adaptations of their past audits. The greatest likeness found in surveys is utilized to catch the most exceedingly terrible spamming conduct.

Most extreme number of audits: In a solitary day posting many surveys reflects anabnormal inspecting design andthis can be utilized as a behavioral component.

Surveying Activity: Reports propose that sentiment spammers are typically not long time individuals from a site. Honest to goodness commentators however utilize their own records from at whatever time to post surveys over a drawn out stretch of time. It is consequently valuable to knw the freshness of a record to recognize spamming.

a.Review Features

i.Extreme Rating: Opinion spamming typically projects entities incorrectly either in a very positive or a very negative light.

ii Rating Deviation: It has been observed that the ratings of spammers deviate from the average ratings given by other genuine reviewers.

iii Early Time frame: Spammers often review early to inflict spam as the early reviews can greatly impact the people’s sentiment on the entity.

We propose another spam identification procedure utilizing the content bunching in view of vector space show. Our strategy processes disjoint bunches naturally utilizing a circular k-implies calculation for all spam/non-spam surveys and acquires centroid vectors of the groups

for removing the group depiction. For every centroid vectors, the name ('spam' or 'non-spam') is doled out by computing the quantity of spam survey in the bunch. At the point when new audit arrives, the cosine similitude between the new survey vector and centroid vector is ascertained. At long last, the mark of the most significant group is doled out to the new audit.

3.1.5 FEATURE SELECTION

A large feature set usually contains a number of irrelevant or redundant features, which may hide useful information from the relevant features . In order to deal with this problem, feature selection is so proposed to select a small relevant feature subset by removing irrelevant and redundant features. It is expected that feature selection can shorten the training time and improve the classifications performance by using all features.

Feature selection has two main objectives, which are to minimise the number of selected features and maximise the classification accuracy. Therefore, the following minimisation fitness function is used:

$$\text{fitnessFS} = \alpha * \text{ErrorRate} + (1 - \alpha) * \#\text{selected}/\#\text{all}$$

where ErrorRate means the classification error rate of the selected features, #selected represents the number of selected features and #all is the total number of original features. α is used to control the contributions of the classification performance and the number of selected features.

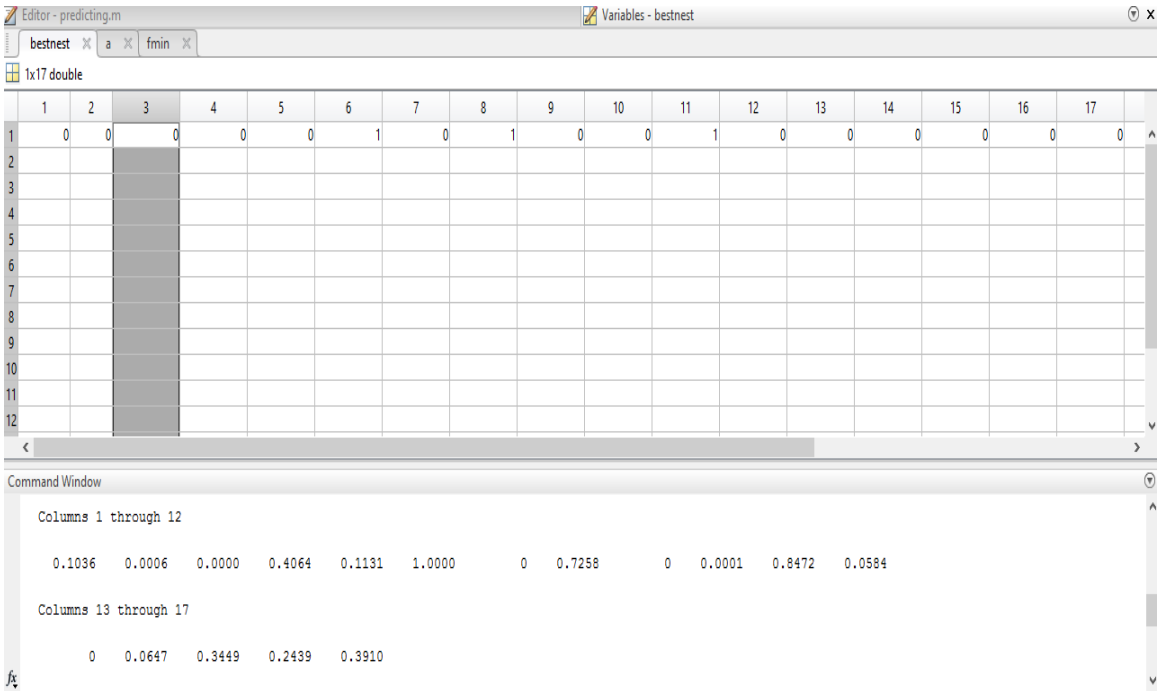

Fig 3.1.features selected using cuckoo search algorithm

3.2 FUNCTIONAL REQUIREMENTS

Following are the modules figured:

i.User module.

ii. Administrators module.

The functionality of these modules is:

User module: The user will input a review to check whether it is spam or not.

Administrator module: The administrator will provide training data sets and monitor the learning process.

The role of Administrator is:

- Provide training data set
- Validate data
- Monitor training process

The role of User is:

- Provide input for analysis
- Review input
- Provide feedback for further scope of improvement

3.3 NON-FUNCTIONAL REQUIREMENTS

Reliability of system and availability requirements

The product shall be available for use 24 hours a day, 365 days a year.

Robustness and Fault Tolerance requirements

At whatever point it loses its connection to the focal server the item ought to be sufficiently competent to keep on operating in nearby module.

Scalability requirement

The product must be capable enough of processing the existing number of customers.

Longevity requirements

Within the Largest maintenance budgets for minimum of few years it is expected that the product shall be expected to operate for that time being.

Maintainability and Support requirements

Reports need be available at any time of requirement.

Installation Requirements:

Wellbeing Critical necessities

Portrayal of the apparent hazard, calculates that could harm.

Execution prerequisites

The reaction might be sufficiently quick to abstain from interfering with the client's stream of thought.

Usability

Easy to use.

Security

Authentication of the user, access control, data integrity, and data privacy.

Supportability

Must be compatible with all major desktop browsers and mobile devices.

3.3 DESIGN DIAGRAMS

3.3.1 USE CASE DIAGRAM

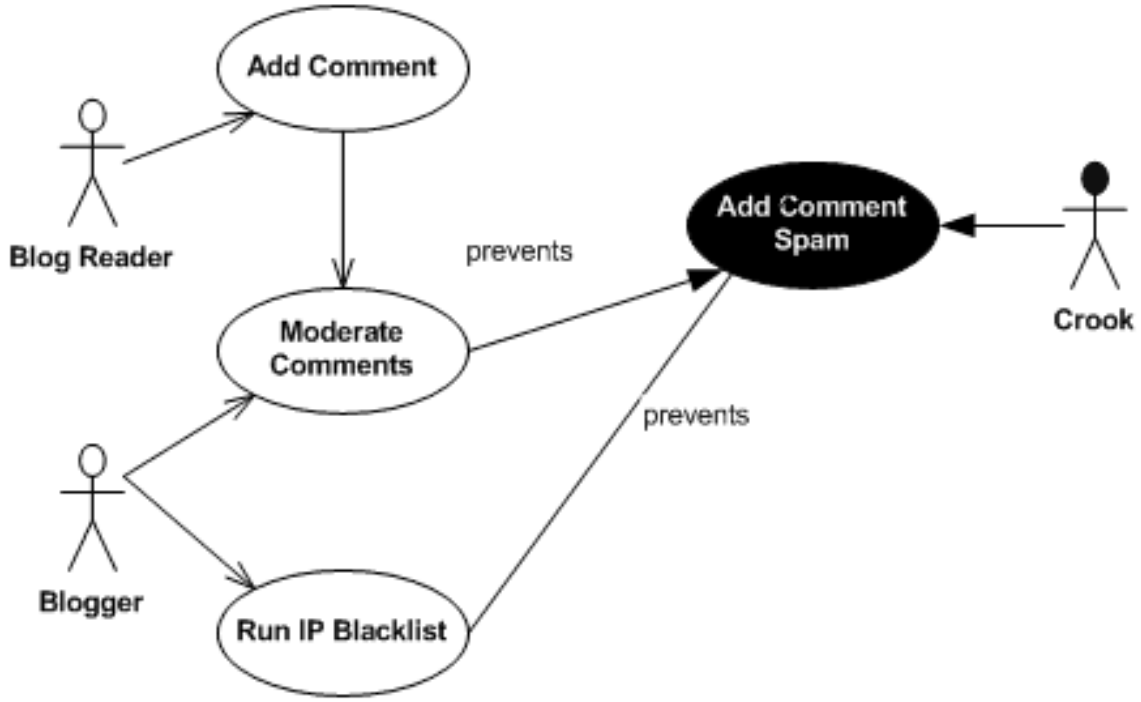

Fig 3.2. Use case diagram

3.3.2 CONTROL FLOW DIAGRAMS

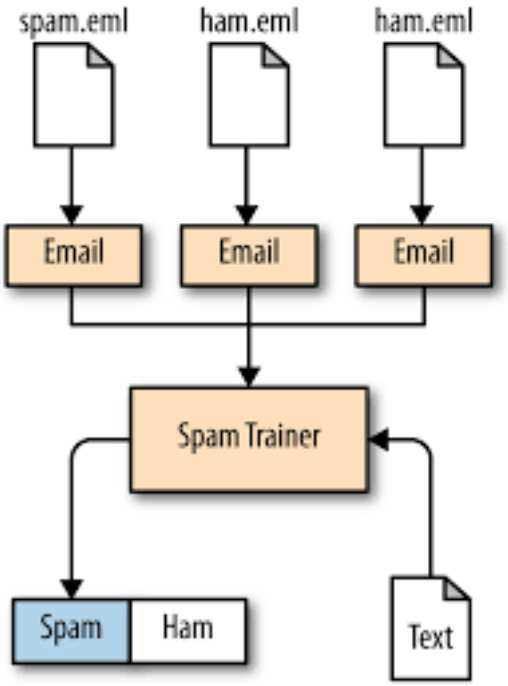

Fig 3.3. spam

CHAPTER 4: IMPLEMENTATION DETAILS AND ISSUES

4.1 IMPLEMENTATION DETAILS AND ISSUES

4.1.1 IMPLEMENTATION ISSUES

We have implemented Waterfall Method. We have gathered all our requirements like training set, testing set, algorithms that we will be using, and then we proceeded making our project dividing it into phases and checking the working of each phase and then proceeding onto the next phase. We have worked together, checked each other's work and provided ourselves with collective ownership. First we have cleaned our dataset by preprocessing ,after that we have calculated the features and further we have used clustering algorithm (cuckoo search) for implementation of our project.

This Linguistic components investigation works averagely . In spite of the fact that, we watch that this investigation is practically identical to the arrangement done physically by human annotators in grouping the same dataset. It was found that people have a precision level of under 60% for the same dataset characterization undertaking. At the point when numerous gatherings were made a request to arrange the dataset, their simultaneousness of results was quite low. Hence, in tricky surveys recognition our etymological elements model is tuned in to the human instinct.

Really it is watch that the spammers utilize an arrangement of words which arealmost in contrast with the certified audit journalists. This property is sufficiently useful for us to order spam conduct. Our underlying theory is additionally demonstrated.

We find that spammers and non-spammers may have utilized comparative words, however the recurrence of its use from the word-sets has an enormous effect.

We can utilize the N-gram display when all is said in done in a wide range of situations, not to mention lodging surveys as the essential thought stays same and this strategy functions admirably on a wide range of datasets.

4.1.2 ALGORITHM

- **Feature Extraction:-** New attributes are created using linear combinations of original attribute.
- **Association Rules:-** The items that can co-occur in the data it helps to find them and specifies some rules that control their co-occurrence.
- **Anomaly Detection:-** It Identifies items that do not satisfy the characteristics of normal data.

SUPPORT VECTOR MACHINE

In machine taking in, the support vector machines which also bolster vector networks are related with learning calculations administered learning models. SVM break down the information utilized for order and than relapse examination. Given an arrangement of preparing illustrations, each set apart as having a place with either of two classifications, a SVM preparing calculation manufactures a model that allocates new cases to one classification or the other, making it a non-probabilistic double direct classifier. A SVM model is a portrayal of the cases as focuses in space, mapped so that the cases of the different classes are partitioned by an unmistakable hole that is as wide as could reasonably be expected. New cases are then mapped into that same space and anticipated to have a place with a class in light of which side of the hole they fall on. At the point when information are not named, directed learning is unrealistic, and an unsupervised learning methodology is required, which endeavors to discover regular bunching of the information to gatherings, and after that guide new information to these shaped gatherings. The bunching calculation which gives a change to the bolster vector machines is called support vector grouping and is frequently utilized as a part of modern applications either when information is not marked or when just a few information is named as a preprocessing for an arrangement pass.

NATURE INSPIRED ALGORITHM : CUCKOO SEARCH

Cuckoo seek algorithmis a nature-enlivened populace based algorithm.It was initially proposed by Yang and Deb in 2009 to take care of enhancement issues . The CS calculation is enlivened by the brood parasitism of cuckoo species that lay their eggs in the homes of the host winged

animals of different species with the point of getting away from the guardians interest in raising their posterity. The methodology is likewise valuable for limiting the danger of egg misfortune to different species, as the cuckoos can appropriate their eggs among various diverse homes. now and again it might happens that the host winged creatures find the outsider eggs in their homes. In such cases, the host flying creature can take diverse responsive activities shifting from discarding such eggs to just leaving the home and construct another one somewhere else. Notwithstanding, the parasites have at their turn created modern techniques, (for example, shorter egg brooding periods, fast settling development, and egg tinge or example emulating their hosts) to guarantee that the host flying creatures will take look after the homes of their parasites.

4.2 RISK ANALYSIS AND MITIGATION

Table 4.1. Risk Analysis and mitigations

Risk	Description Of the Risk	Areas of The risk found	Probability	Impact	RE	Risk Selected for Mitigation	Mitigation Plan	Contingency Plan
1	Training wit wrong datas	Data Training	0.5	5 (High)	2.5	Yes	Validate the dataset before providing it fo training.	NA
2	Dataset not sufficiently large	Data Training	0.3	5 (High)	1.5	Yes	Provide optimal datas for training	NA
3	Error in Cod	Execution	0.4	3 (Medium	1.2	Yes	Debug and Recompile th code	NA
4	Invalid Inpu from user	Runtime	0.5	1 (Low)	0.5	No	NA	Display Input not valid

CHAPTER 5 : TESTING

5.1 TESTING PLAN

Test Plan for this application consist different areas of testing mentioned as follows**:**

Table 5.1 Testing plan

Types of test performed	**Will the test be performed**	**Comments and explanations**	**Software components used**
Requirement testing	Yes	Performed our code in python and matlab	Python Matlab
Unit	Yes	Few points tested and verified with results in form of excel sheets	Python Matlab
Integration	No	Have used two different languages ,hence implementation is different for both softwares	Python Matlab
Performance	Yes	Results are accurate	Python ,matlab
Security	No	No security functions are not implemented	
Volume	Yes	Volume of data is limited as yelp dataset is used	Excel
Load	Yes	Code lines limited	
Support vector Machine	Yes	For Machine learning	Matlab
Cuckoo Algorithm	Yes	Algorithm runs for Multiple itterations	Matlab

5.2. DECOMPOSITION OF COMPONENTS AND THE TYPE OF TESTING REQUIRED

Table 5.2. Type of Testing

s.no	List of vario components th require testing	Type of testi required	Technique for testing
1.	Data scraping, Data cleaning, Sentiment analysis	1.1 Requirements Testing 1.2 Unit Testing 1.3 Integration Testing	**Research ,black box testing**
2.	Data scraping, Data cleaning, Sentiment analysis	2.1 Performance Testir 2.2 Load Testing 2.3 Volume Testing	**White box testing**

5.3 TEST CASES

Table 5.3. Test cases

TEST CASE ID	INPUT	EXPECTED OUTPUT	STATUS
1.1	Conforming to previous researches and technical documents	Correct understanding of requirements	Pass
2.1	Applying machine learning algorithm	Providing result spam or no spam for reviews	Pass
1.2	Extracted dataset	Punctuation and other special character and phrases removed	Pass
1.2	Cleaned dataset	Segregation of positive and negative reviews	Pass
2.2	Features extraction	Extracted required features from the dataset ar stored in excel file	Pass
2.2	Dataset reviews	Fast extraction and segregation	Pass
2.2	Distorted reviews	segregation	Fail
2.3	Mixed set of dataset	Segregated reviews	Fail
2.1	Accuracy	Same or better accuracy when done with feature selection technique	Pass
2.3	Large dataset	Storing and segregation	Pass

5.4 LIMITATIONS OF THE SOLUTION

The solution worked out till now is:

- Preparation of architecture.
- Implementation of Data scraping
- Data cleaning
- Implementation of sentiment analysis
- Features extraction
- Cuckoo search
- Feature selection
-
- Applied support vector machine algorithm for machine learning
- Accuracy computation

Some limitations were realized during the project, as listed below:

- Reviews with various special characters are difficult to classify.
- Reviews with only images and characters cannot be classified.
- Review rankings can overlap so it will be difficult for classifying such reviews.
- Data extraction of reviews for some products differ from expected dataset.
- Product with few reviews are difficult to classify and accuracy also decreases in detection of authenticated product.
- Results are not in same order as the reviews in database

CHAPTER 6: FINDINGS& CONCLUSION

6.1 FINDINGS AND CONCLUSION

The cuckoo search analysis works pretty effectively and also provides reasonably good results in detecting the spam reviews. We observe that POS model as well as linguistic features and further more features provide a secondary support for our classification model. The combined model gives better results as it encompasses the psychological tendency of spammer. Also, we understood that the without the help of reviewers and their information this analysis is incomplete. Sentiments is also being incorporated in the model that provide reasonably good results. Destinations like Amazon, have as of late presented an alternative that imprints confirmed purchaser against the audits, along these lines taking a jump in keeping away from the effect of supposition spam. The audits of these checked purchasers can be utilized as a benchmark to divide the genuine surveys from the fake ones.

6.3 FUTURE WORK

Some of the user metadata such as written reviews number ,timeframe of writing the reviews, IP address of the reviewer, etc. could be very crucial for our spam analysis and could help in determining fraudulent reviews and spams.

Lamentably, because of security concerns, we don't get the client data on the said sites and just those sites can break down the client information inside. We could likewise check for the veritable remainder of the content by coordinating the surveys with data accessible in the official sites for the given items, for example, hardware audits could be checked against en device or tech crunch and inn surveys could be checked against basic commentators for the same. In any case, we could utilize this proposed function as a basline for further enhancements in this examination area.we would attempt to add more components to our venture for much better precision.

References

Book:

[1] Ghahramani, Zoubin. "Unsupervised learning." *Advanced lectures on machin learning*.Springer Berlin Heidelberg, 2004.

[2] Han, Jiawei, Jian Pei, and MichelineKamber. *Data mining: concepts and techniques*, Elsevier, 2011.

[3] Yang, Xin-She. *Nature-inspired metaheuristic algorithms*, Luniver press, 2010.

[4] Rout, Jitendra Kumar, "Deceptive review detection using labeled and unlabeled data." *Multimedia Tools and Applications* , 2016.

Online:

[1] http://in.mathworks.com/discovery/unsupervised-learning.html

[2] https://en.wikipedia.org/wiki/K-means_clustering

[3]https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0029-9#CR11

[4] http://www.slideshare.net/dalmiaayushi/sentiment-analysis-in-twitter

[5] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

[6] http://web.stanford.edu/class/cs345a/slides/12-clustering.pdf

[7] http://www.ijettcs.org/Volume4Issue1/IJETTCS-2015-01-22-43.pdf

[8]http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[9]http://www.cs.uic.edu/~liub/publications/cikm-2010-final-spam.pdf