Spam Detection in Social Bookmarking Systems

Prepared By

Mittal Sejpal

12MICT52

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2014

Spam Detection in Social Bookmarking Systems

Major Project

Submitted in partial fulfillment of the requirements

For the degree of

Master of Technology in Computer Science and Engineering(Networking Technologies)

Prepared By

Mittal Sejpal

(12MICT52)

Guided By

Prof. Priyank Thakkar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2014

Certificate

This is to certify that the Major Project Report entitled "**"Spam Detection in Social Bookmarking Systems"** submitted by **Mittal Sejpal (Roll No: 12MICT52)**, towards the partial fulfillment of the requirements for the degree of Master of Technology in Computer Science and Engineering(Networking Technologies) of Nirma University, Ahmedabad is the record of work carried out by her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Priyank Thakkar
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof. Gaurang Raval
Assistant Professor
Coordinator M.Tech - CSE
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr K Kotecha
Director,
Institute of Technology,
Nirma University, Ahmedabad

Undertaking for Originality of the Work

I, **Mittal Sejpal**, Roll. No. **12MICT52**, give undertaking that the Major Project entitled "**Spam Detection in Social Bookmarking Systems**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering(Networking Technologies)** of Nirma University, Ahmedabad, is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Prof. Priyank Thakkar

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Priyank Thakkar**, Assistant Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

My deepest thank you is extended to **Prof. Gaurang Raval**, PG CSE(NT) - Coordinator, Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad for an exceptional support and continual encouragement throughout the Major Project.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr K Kotecha**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, and Ahmedabad for their special attention and suggestions towards the project work.

The blessings of God and family members make the way for completion of Project. I am very much grateful to them.

- Mittal Sejpal

12MICT52

Abstract

Social bookmarking websites have recently become well-known for collecting and sharing of interesting Web sites among users. People can add Web pages to such sites as bookmarks and allow themselves as well as others to work on them. One of the key features of the social book marking sites is the ability of annotating a Web page when it is being bookmarked. The annotation usually contains a set of words or phrases, which are collectively known as tags that could reveal the semantics of the annotated Web page. Efficient and effective search of Web pages can then be achieved via such tags. However, spam tags that are irrelevant to the content of Web pages often appear to deceive other users for malicious or commercial purposes. Manual Spam Detection is very Difficult. The main purpose is to automate the manual spam Detection.

In this work, focus is on the detection of spam users in Social Bookmarking Systems. Experimental evaluation is done using ECML PKDD discovery challenge 2008 dataset. Nave bayes and K Nearest Neighbour classifier are applied on all three Information Retrieval Models(Boolean, Word Count and TF-IDF).

Information Gain is used as feature selection measure and further all the Information Retrieval Models are trained with the mentioned classifiers. Naive Bayes Classifier gives Promising results with only few attributes with feature selection.

Contents

List of Figures

Chapter 1

Introduction

1.1 General

A Social Bookmarking Service is an incorporated online service which empowers users to include, expound, alter, and offer bookmarks of web documents. Many online bookmark administration services have propelled since 1996; Delicious, established in 2003, promoted the expressions "social bookmarking" and "tagging". Tagging is a critical characteristic of social bookmarking Systems, empowering clients to sort out their bookmarks in adaptable ways and create imparted vocabularies known as folksonomies.

Figure 1.1: Social Bookmarking Websites

Not at all like file sharing, social bookmarking does not spare the resources themselves, just bookmarks that reference them, i.e. a connection to the bookmarked page. Portrayals may be added to these bookmarks as metadata, so users may comprehend the substance of the asset without first expecting to download it for themselves. Such portrayals may be free content remarks, votes on the side of or against its quality, or labels that all in all or collectively turn into a folksonomy. Folksonomy is additionally called social tagging, "the procedure by which numerous clients include metadata as pivotal words to imparted substance".

Most social bookmark services encourage users to compose their bookmarks with casual tags rather than the conventional program based arrangement of organizers, albeit a few services characteristic classifications/envelopes or a consolidation of envelopes and tags. They additionally empower seeing bookmarks connected with a picked tag, and incorporate data about the amount of clients who have bookmarked them. Some social bookmarking services likewise draw inductions from the relationship of tags to make bunches of tags or bookmarks.

Numerous social bookmarking services give web nourishes to their arrangements of bookmarks, including records composed by labels. This permits supporters of get mindful of new bookmarks as they are spared, imparted, and tagged by different users. It additionally serves to push your locales by systems administration with other social book markers and teaming up with one another.

As these services have developed and become more prominent, they have included additional characteristics, for example, appraisals and remarks on bookmarks, the capacity to import and fare bookmarks from programs, messaging of bookmarks, web annotation, and gatherings or other interpersonal organization characteristics.

Social Bookmarking Systems have picked up high notoriety now a days. With this developing notoriety the Spam users abuses them as a play area for their exercises. There are two principle objectives of Spam client while setting the connections: Attracting the individuals to their destinations and expanding the page rank of their sites. Regular spam counter-measures, for example, captchas don't sufficiently keep spammers from abusing

the framework[1].

In Social bookmarking sites users can rather store and access their bookmarks online through a Web interface. The put away data is sharable among users, considering enhanced looking. Any framework that is profoundly reliant on client produced substance is powerless against spam in one structure or an alternate[2].

Spam Users get spurred for promoting and progression toward oneself of their sites. Social Bookmarking Websites give a huge and constantly developing pool of pool of potential clients. Actually a spam post is more alluring than a non-spam post[3].

Spamming in Social Bookmarking Systems has become a crucial challenge affecting both users and service providers. Users face spamming obstacles while doing activities like web based searching since spam users uses techniques or keywords to increase the page rank of their websites[4].

The main challenge is the characteristics and behaviour of spam user's change over time hence maintaining the rules becomes a difficult task. It is very difficult to manually detect spam users because of huge number of user's data etc [4].

1.2 Definition

The information structure that support a labeling systems is a group turned obsolescent known as "folksonomy," formally spoke to as a hyper-chart. In this point of view, centers incorporate clients, possessions and names, and each annotation incorporates a hyper-edge to the chart, joining a client, the elucidated holding, and the picked tag. Since every client can without much of a stretch add to the folksonomy, the structure of the diagram is through and through client driven, and a noxious client can misuse this control to make some substance more unmistakable, drive client action to picked targets, and with everything taken into account to sully the folksonomy. We insinuate these sorts of abuses of communitarian annotation schemas as social spam. Perceiving social spam regularly

and adequately is a key test in making social annotations suitable for any given system and for the web free to move around at will.

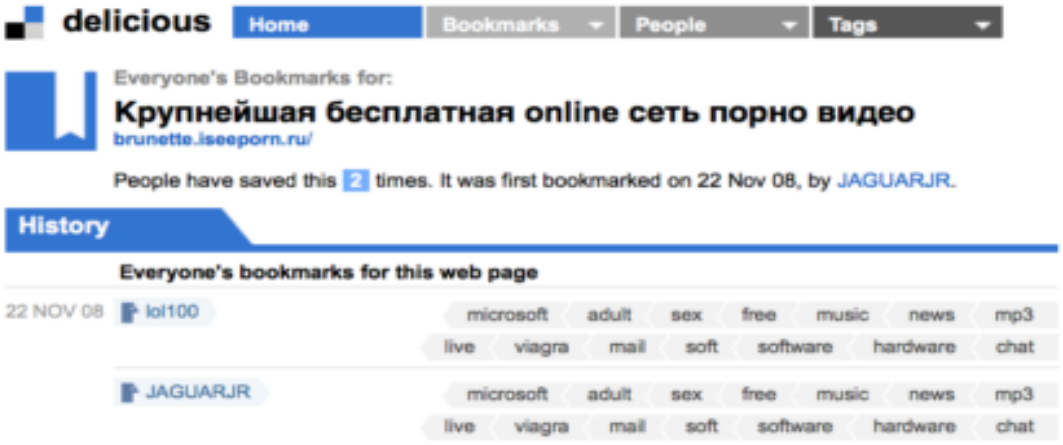

Figure 1.2: A spammer using two characters posts a Russian porn site on delicious.com. The spammer uses standard marks, for instance, "music," "news" and "programming," which are immaterial to each one in turn and the site.

The spam users use the non spam keywords or tags for bookmarking their spam website. Hence Spam Detection in a Social Bookmarking system is a key challenge because Spammers often change their behavoiur thus it becomes vey difficult to identify spam users based on behavoiur.

Chapter 2

Literature Survey

Social spam is a tolerably new research region and the composition is still meager. After a formal importance of the information structures underlying a folksonomy and some establishment on social labeling schemas, we give a brief audit of related work in not well arranged information recuperation and the late development of attention to social spam[5].

The different research papers identified with Spam Detection in Social Bookmarking Systems was done throughout the dissertation stage. The essential highlights of a percentage of the research papers and overview has been said in this part. The points of interest are as said underneath.

2.1 Papers Studied

"Using Language Models for Spam Detection in Social Bookmarking[2]"

- This paper proposed methodology to the spam detection by the utilization of language models. The proposed methodology is focused around the natural thought that comparable users and posts have a tendency to utilize the same language.
- New spam users in the framework are identified by first positioning all the old users in the framework by the KL-divergence of the language models of their posts independently and joined into user profiles and the language model of the new user or post.

- The methodology is examined different things with matching language models at two separate levels of granularity and found that, all in all, matching at the user-level gave the best comes about.
- A conceivable limitation of the methodology in the meantime: spammers will change their conduct about whether and may have done so in the time period the test set begins from.

"A novel supervised learning algorithm and its use for Spam Detection in Social Book marking Systems[3]"

- This paper proposed a novel fast and accurate supervised learning algorithm as a general text classification algorithm for linearly separated data.
- Svms have demonstrated best execution in Text Categorization prepare yet their just inconvenience is the memory necessities and training multifaceted nature.
- The proposed Algorithms take the focal point to progressively refine the beginning classifier. This refinement takes the manifestation of an iterative Rocchio-like relevance feedback - learning system to alter the centroid vectors of the classifications, with a specific end goal to expand the execution of the classifier.
- Every user was distinguished by a remarkable id and after this venture for every user id we had a text fragment speaking to the greater part of his posts.

"Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Book marking[6]"

- The Paper proposed the machine learning method to mechanize the manual spam detection.
- At first the set of Relevant Features i.e. number of posts and number of posted Tags is concentrated from the training data. The concentrated tags are then sorted by their mutual information.
- The tags having high mutual information worth are utilized within the test data and are picked for the classification undertaking.

- In this paper, naive Bayes classifier with fluctuating amounts of chose features were gained from a subset of the given training dataset and assessed on a different acceptance set for discovering the ideal parameter setting.
- The proposed feature selection method gave guaranteeing brings about Detecting spammers.
- Future work might be to consolidate the interrelationship between tags into the methodology for more improved classification execution.

"Using Co-occurrence of Tags and Resources to Identify Spammers[7]"

- This paper proposed an algorithm to recognize spammers from the collaborating frameworks by utilizing a spam score proliferating technique.
- Firstly a graph is built which demonstrates users as hubs and three sorts of relationship between users as edges.
- Particularly, the paper recognized the accompanying sorts of relationship between users: normal tags supplied by users, regular resources clarified by users and basic tag-resource sets utilized by users.
- A set of seed nodes are then chosen whose comparing users are physically surveyed as spammers or not. The personality of the remaining nodes/users is figured by spreading the status of seed nodes through the graph.

"Combining Clustering with Classification for Spam Detection in Social Book marking Systems[8]"

- This paper addresses the issue of learning to classify texts by exploiting data inferred from both training and testing sets.
- In this paper clustering is utilized as an integral venture to text classification, and is connected to the training set as well as to the testing set.
- The algorithm comprises of the accompanying three steps:
    1. Clustering step: The training and testing set both are clustered in this step.

2. Expansion step: The dataset is supplimented with meta-features emerged from the clustering step.
3. Classification step: In this step classifier is trained with the elaborated dataset.

- On all analyses directed, the clustering methodology joined with a SVM/ TSVM classifier indicated improvements over the utilization of a standard SVM/TSVM classifier on its own.
- One constraint of the algorithm is that when another user makes his first post, the same method of clustering, meta-feature expansion, and classification, ought to be connected again for the entire dataset.

"Using Semantic Features to Detect Spamming in Social Book marking Systems[4]"

- The Paper examines potential features that depict the framework's users and outline how we can utilize those features within request to focus spammers through different machine learning models.
- The features used in the paper are those stated in [9].
- The features were then experimented with five Models:
  1. K-Nearest Neighbour Regression.
  2. Gaussian Processes.
  3. Support Vector Machine (SVM).
  4. Neural Networks.
  5. Both SVM and Neural Networks.
- The results affirm the theory that urls could be recognized as a genuine pointer of spam users, although tags might be tricky.

"Social Spam Detection[5]"

- The Paper proposed and investigated six different features that address different properties of social spam.
- These features are then utilized as a part of different machine learning algorithms for classification.

- The ensuing classifiers accomplish accuracy over 98% while keeping up a false positive rate close to 2%.
- The features that were used are as following:
    1. TagSpam.
    2. Plagiarism.
    3. ValidLinks.
    4. TagBlur.
    5. DomFP.
- Combining each of the six features gave guaranteeing results.

Chapter 3

Experimental Evaluation

3.1 Data set

ECML PKDD Discovery Challenge Dataset (spam detection) 2008
The Dataset Contains the Following Files:

1. tas_spam :- Tag assignments: Table That indicates or states who has anotated which tag to which Content.
2. bookmark_spam :- Table that contains Dimension for Bookmark data.
3. bibtex_spam :- Table that contains Dimension for Bibtex data.
4. user_spam :- Table that contains Dimension for user data.

The Following gives the details of the No. of Records in the given Files:

1. tas_spam 2,743,743
2. bookmark_spam 200,094
3. bibtex_spam 152,906
4. user_spam 7171

After processing the above given files in the dataset as per our requiremnts we have genrated the dataset with a file containg 7171 usere and 71911 unique tags.

3.2 Tools

- **Net Beans IDE 7.3.1**

Netbeans is a coordinated nature's turf (IDE) for creating essential with Java, additionally with different dialects, specifically PHP, C/C++, and Html5.It is likewise a requisition stage structure for Java desktop requisitions and others.

The Netbeans IDE is composed in Java and can run on Windows, OS X, Linux, Solaris and different stages supporting a perfect JVM.

The Netbeans Platform permits provisions to be created from a set of measured programming segments called modules. Requisitions focused around the Netbeans Platform (counting the Netbeans IDE itself) could be stretched out by outsider developers.

Structure for disentangling the advancement of Java Swing desktop provisions. The Netbeans IDE pack for Java SE holds what is required to begin creating Netbeans plugins and Netbeans Platform based provisions; no extra SDK is needed.

Requisitions can introduce modules powerfully. Any requisition can incorporate the Update Center module to permit clients of the provision to download digitally marked redesigns and new characteristics straightforwardly into the running provision. Reinstalling an overhaul or another discharge does not drive clients to download the whole requisition once more.

The stage offers reusable administrations regular to desktop provisions, permitting designers to concentrate on the rationale particular to their requisition. Among the characteristics of the stage are:

- User interface administration (e.g. menus and toolbars).
- User settings administration.
- Storage administration (sparing and stacking any sort of information).

- Window administration.
- Wizard schema (backings orderly dialogs).
- Netbeans Visual Library.
- Integrated improvement apparatuses.

Netbeans IDE is a free, open-source, cross-stage IDE with inherent backing for Java Programming Language.

Usage: NetBeans is used for implementing supervised method Naive Bayes on the dataset.

- **Weka 3.7**

Weka (Waikato Environment for Knowledge Analysis) is a gathering of machine Learning calculations for information mining assignments. The calculations can either be connected specifically to a dataset or called from your Java code. Weka holds apparatuses for information pre-processing, order, relapse, bunching, cooperation principles, and visualization. It is additionally appropriate for creating new machine taking in plans.

The Weka (attested Weh-Kuh) workbench holds a social affair of visualization instruments and estimations for data examination and prescient exhibiting, ...together with graphical customer interfaces for straightforward access to this convenience. The primary non-Java manifestation of Weka was a TCL/TK front-end to (essentially outcast) exhibiting estimations realized in other programming lingos, notwithstanding data preprocessing utilities in C, and a Makefile-based skeleton for running machine taking in tests. This one of a kind structure was generally arranged as a mechanical assembly for dismembering data from agrarian areas, however the later totally Java-based adjustment (Weka 3), for which progression started in 1997, is as of now used as a piece of various unique demand locales, particularly for informational purposes and examination.

Preferences of Weka include:

- Free accessibility under the GNU General Public License.

- Convenientce, since it is completely executed in the Java programming dialect and in this manner runs on just about any advanced registering stage.
- A complete accumulation of information preprocessing and demonstrating procedures.
- Convenience because of its graphical client interfaces.

Usage: Weka is especially used for pre-processing of the dataset.

3.3 Pre-Processing

Pre-processing of dataset is necessary to improve the quality of data thereby helping to improve subsequent classification processes. Pre-processing involves following steps:
All the characters are changed over to lower case and prevent words are expelled from each of the tags. Each of the tag is then characterized as vector in multidimensional Euclidean space. Make diverse representation of each one tag utilizing Wekas String - Towordvector device. The axes of this multidimensional Euclidean space are terms showing up in tag accumulation. Three separate representations in particular Boolean, sack of-words and TFIDF of these vectors are practiced in this study.

In Boolean representation, these vectors are Boolean vectors and every component of the vector speaks to nonappearance or vicinity of the comparing term in the relating tag.

Every component of the vector, if there should arise an occurrence of pack of-words representation is a common number showing how frequently the comparing term has showed up in the relating tag.
These vectors are genuine esteemed vectors when spoken to as far as TFIDF qualities of the terms. TF, IDF and TFIDF are defined in mathematical statements (3.1),(3.2) and (3.3) individually.

As the term infers, TF-IDF figures values for each one word in a document through a backwards extent of the frequency of the word in a specific document to the rate of documents the word shows up in.

$$TF = \log(1 + f_{ij}) \tag{3.1}$$

$$IDF = f_{ij} * \log(\frac{numberoftags}{numberoftagswithwordI}) \tag{3.2}$$

$$IFIDF = TF * IDF \tag{3.3}$$

FLOWCHART:

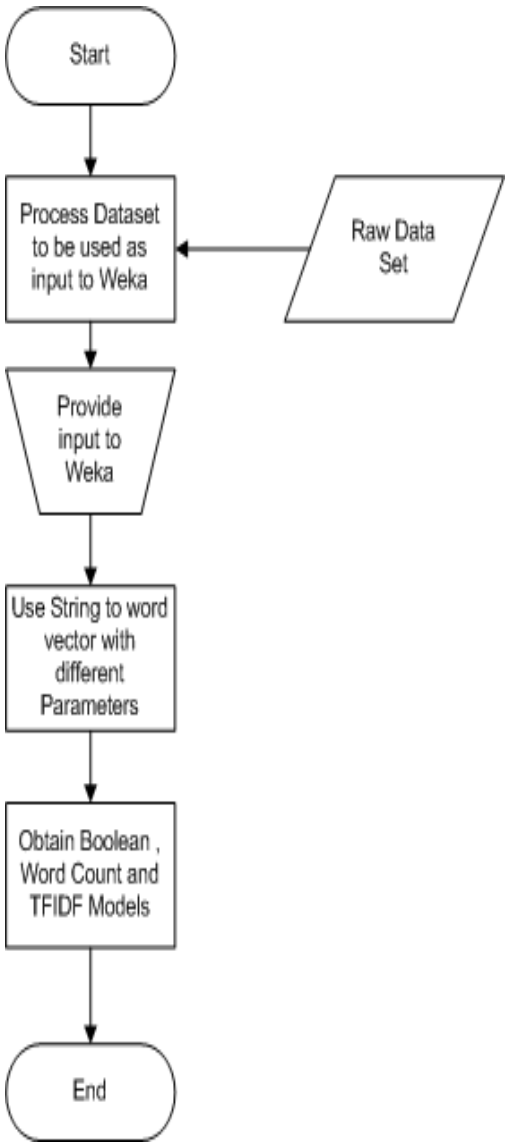

Figure 3.1: Flow chart of Pre-processing

3.4 Evaluation Measures

The parameters for evaluating the performance of the classification algorithms are Accuracy, Precision, Recall and F - Measure. Accuracy of a classifier on a given test set is the percentage of test set items that are correctly classified by the classifier. In this Paper Non - Spammer user is considered as positive class and Spammer user is considered as negative class. Accordingly true positive (TP), false negative (FN), false positive (FP) and true negative (TN) are defined as under[10].

TP (True Positives):- Refers to the number of positive items that were correctly labelled by the classifier.
FP (False Positives):- Refers to the number of negative items that were incorrectly labelled as positive.
TN (True Negatives):- Refers to the number of negative items that were correctly labelled by the classifier.
FN (False Negatives):- Refers to the number of positive items that were mislabelled as negative.

The confusion matrix and the equations of Precision and Recall based on the above interpretations are as follows:

Confusion Matrix:

	P' (Predicted)	N' (Predicted)
P (Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative

$$Precision_{positive} = \frac{TP}{TP + FP} \tag{3.4}$$

$$Precision_{negative} = \frac{TN}{TN + FN} \tag{3.5}$$

$$Recal_{positive} = \frac{TP}{TP + FN} \tag{3.6}$$

$$Recal_{negative} = \frac{TN}{TN + FP} \tag{3.7}$$

Precision is defined as the weighted average of precision positive and negative while Recall is defined the weighted average of recall positive and negative. F - Measure is calculated using the following equation:

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.8}$$

The accuracy of a classifier on a given test set is the percentage of test set items that are correctly classified by the classifier. The Formula of the Accuracy is given by the following equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.9}$$

3.5 Classification Methods

The task of catching Spam user in Social Bookmarking System is demonstrated as binary classification issue in this study. The two classes acknowledged are Spammer and truthful Non - Spammer. Two well-known machine learning methods in particular naive Bayes and K Nearest Neighbor are utilized to learn the classifier.

The algorithms uses K fold Stratified. Here k is taken as 5.

Naive Bayes Method

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Supervised learning can be naturally studied from a probabilistic point of view. The task of classification can be regarded as estimating the class posterior probabilities given a test example.

Naive-Bayes classifier accepts class conditional independence. Given test data Bayesian classifier predicts the probability of data having a place with a specific class. To anticipate probability it utilizes idea of Bayes' theorem. Bayes' theorem is valuable in that it gives a method for computing the posterior probability, $P(C|X)$, from P(C), $P(X|C)$, and P(X). Bayes' theorem states that

$$P(X|C) = \frac{P(X|C)P(C)}{P(X)} \tag{3.10}$$

Here $P(C|X)$ is the posterior probability which lets us know the probability of theory C being true given that event X has happened. For our situation theory C is the probability of having a place with class Spammer or Non - Spammer and event X is our test data. $P(X|C)$ is a conditional probability of event of event X given speculation C is true. It could be evaluated from the training data. The working of naive Bayesian classifier, or basic Bayesian classifier, is outlined as follows:

Let m classes C1, C2..., Cm and event of occurrence of test data, X, is given. Bayesian classifier classifies the test data into a class with most elevated probability. By Bayes'

theorem (Equation 3.10),

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{3.11}$$

Given data sets with numerous attributes (A1, A2, ...,an), it might be it would be extremely computationally expensive to calculate $P(X|Ci)$. o as to decrease calculation in assessing $P(x|Ci)$, the naive suspicion of class conditional independence is made. This presumes that the qualities of the attributes are conditionally free of each other, given the class label of the tuple (i.e. that there are no reliance connections among the attributes).therefore,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{3.12}$$

$$= P(X_1|C_i) \times P(X_2|C_i) \times ... \times P(X_n|C_i) \tag{3.13}$$

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

K Nearest Neighbor

K Nearest Neighbor (KNN) is an algorithm that is extremely easy to see yet works extraordinarily well in practice. Likewise it is shockingly adaptable and its provisions range from vision to proteins to computational geometry to diagrams thus on . Most individuals take in the calculation and don't utilize it much which is a compassion as an astute utilization of KNN can make things extremely basic. It additionally may amaze numerous to realize that KNN is one of the main 10 information mining algorithms.

KNN is a non parametric lethargic taking in calculation. That is a really succinct proclamation. When you say a method is non parametric , it implies that it doesn't make any presumptions on the underlying information conveyance. This is really helpful , as in this present reality , the vast majority of the reasonable information does not comply with the common hypothetical suspicions made (eg gaussian mixtures, straightly distinguishable and so forth) . Non parametric algorithms like KNN act the hero here.

It is likewise a lethargic calculation. This means it doesn't utilize the preparation information focuses to do any generalization. At the end of the day, there is no unequivocal preparing stage or it is extremely insignificant. This methods the preparation stage is really quick . Absence of generalization implies that KNN keeps all the preparation information. All the more precisely, all the preparation information is required throughout the testing stage. (Well this is an embellishment, however not a long way from truth). This is as opposed to different procedures like SVM where you can dispose of all non help vectors without any issue. Most of the languid algorithms particularly KNN settles on choice focused around the whole preparing information.

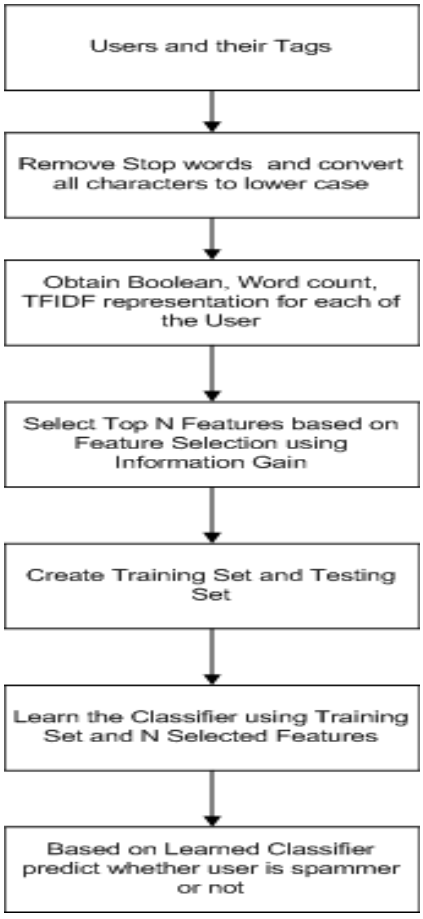

Figure 3.2: Flow Chart of Implementation

3.6 Results And Discussion

3.6.1 Results of Classification Algorithms

Tables I shows the result from all the classification algorithms that were implemented. It is paramount to tell that Naive Bayes and KNN classifiers are utilized as learning algorithm. Weka is used for pre-processing of data and NetBeans is used for implementation of mentioned classification Algorithm.

The following are the results of all the performance measure parameters for all the algorithms:

Result of Two Class (1, 0) **1:** Spammer **0:** Non spammer with Stratified (5 Fold) cross - validation , and No. of Attributes: 71911

5-fold cross validation	Accuracy comparision		
	Boolean	Word Count	Tf-Idf
Naive Base	97.544	78.297	84.395
Knn	97.423	96.931	96.984

Table 3.1: Accuracy Comparision of Classification Algorithms Using all the attributes

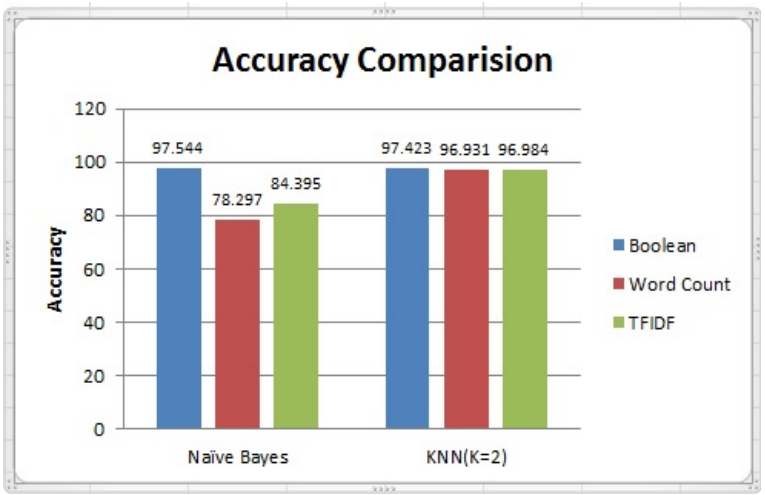

Figure 3.3: Accuracy Comparision Chart

From the Above Chart it can be viewed Boolean Representation of Naive Bayes Classifier works comparatively well with all the attributes.

3.6.2 Results of Feature Selection

Feature selection also known as attribute selection is used to reduce the size of the dataset by removing redundant or irrelevant attributes. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection will reduce the set of terms to be used in classification, thus improving both efficiency and accuracy.

Feature selection is applied on all the dataset and classification methods are applied on 5, 10, 15, 20, 30, 60, 100, 200, 300, 500, 1500, 2000, 3000, 4000, 5000 etc and likewise on all the features of the dataset. The total numbers of features are 71911.

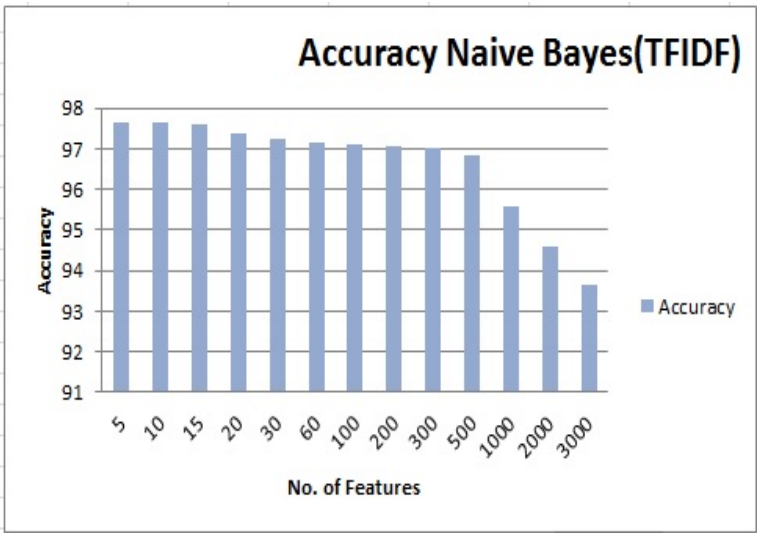

Figure 3.4: Accuracy of TFIDF using Nave Bayes Classifier with diff. no. of Features

From Figure 3.4 it can be viewed that accuracy of 96% is achieved while considering only 500 Attributes using Naive Bayes Classifier which is same as achieved in[6].
Graphical results indicate that recognizing few attributes to assemble classifier with characteristic determination gave same come about as acknowledging all attributes in managed techniques.we additionally infer that after certain point if number of attributes builds that won't effect in expanding Accuracy.

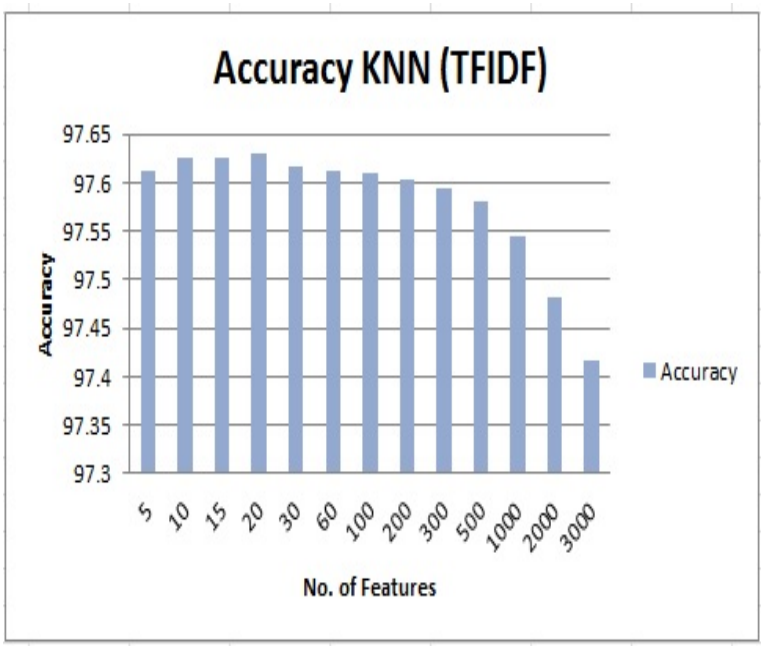

Figure 3.5: Accuracy of TFIDF using KNN with diff. no. of Features

Chapter 4

Conclusion and Future Work

4.1 Conclusion

In this work, Nave Bayes and KNN Classifiers are used with all IR model (Boolean, Word Count, TFIDF).Further Feature Selection is used with Information Gain.
The obtained results signifies that for building the classifiers considering few attributes is equivalent to the results obtained using all attributes. It also indicates after a point the Accuracy of the classifier cannot not be improved. This indicates that accuracy of a classifier can be achieved by relevant attributes not a large number of attributes. Hence Feature selection is used to boost the efficiency of classification. The computation becomes less to construct the model using Feature Selection.

4.2 Future Work

The Accuracy of Nave Bayes classifier starts declining as the number of attributes increases. The future work would be to work on combining the interrelationship between tags with feature selection for better classification performance.

References

[1] I. Workshop, E. Conference, M. Learning, and K. Discovery, "Ecml pkdd discovery challenge 2008," *ECML PKDD Discovery Challenge 2008*, vol. 2008, 2008.

[2] T. Bogers, "Using Language Modeling for Spam Detection in Social Reference Manager Websites," *Using Language Modeling for Spam Detection in Social Reference Manager Websites*, 2009.

[3] A. Gkanogiannis and T. Kalamboukis, "A novel supervised learning algorithm and its use for spam detection in social bookmarking systems," in *In Proc. Europ. Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD*, 2008.

[4] A. Madkour, T. Hefni, A. Hefny, and K. S. Refaat, "Using Semantic Features to Detect Spamming in Social Bookmarking Systems," 2008.

[5] B. Markines, C. Cattuto, and F. Menczer, "Social spam detection," *Social Spam Detection*, vol. 2009, 2009.

[6] C. Kim and K. B. Hwang, "Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking," in *ECML PKDD Discovery Challenge 2008*, 2008.

[7] R. Krestel and L. Chen, "Using Co-occurence of Tags and Resources to Identify Spammers," 2008.

[8] A. Kyriakopoulou and T. Kalamboukis, "Combining clustering with classification for spam detection in social bookmarking systems," 2008.

[9] B. Krause, C. Schmitz, A. Hotho, and G. Stumme, "The anti-social tagger - detecting spam in social bookmarking systems," in *Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web*, 2008.

[10] J. P. Jiawei Han, Micheline Kamber, *Data Mining Concepts and Techniques*. Morgan Kaufmann, 3 ed., July 2011.

[11] S. Young, I. Arel, T. P. Karnowski, and D. Rose, "A Fast and Stable Incremental Clustering Algorithm,"