Opinion Spam Detection with Feature Selection and PU Learning

By

Rinki Patel
11MICT53

NIRMA
UNIVERSITY
INSTITUTE OF TECHNOLOGY

Department of Computer Science & Engineering
Institute of Technology, Nirma University
Ahmedabad

May,2014

Opinion Spam Detection with Feature Selection and PU Learning

Major Project

Submitted in partial fulfillment of the requirements
for the degree of
Master of Technology in Information & Communication Technology

By

Rinki Patel
11MICT53

Guided By

Prof. Priyank B. Thakkar

Department of Computer Science & Engineering
Institute of Technology, Nirma University
Ahmedabad

May, 2014

Certificate

This is to certify that the Project entitled **"Opinion Spam Detection using Feature Selection and PU Learning"** submitted by **Patel Rinki Hitesh (11MICT53)**, towards the submission of the Major Project for requirements for the degree of Master of Technology in Information and Communication Technology at Institute of Technology of Nirma University, Ahmedabad is the record of work carried out by her under our supervision and guidance. In our opinion, the submitted work has reached a level required for being accepted for examination.

Date: Place: Ahmedabad

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

Prof. Gaurang Raval
Associate Professor
Coordinator M.Tech - ICT
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, **Rinki Patel**, Roll. No. **11MICT53**, give undertaking that the Major Project entitled "**Opinion Spam Detection using Feature Selection and PU Learning**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Information & Communication Technology** 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 B Thakkar
(Signature of Guide)

Acknowledgements

The beatitude, bliss & euphoria that accompany the successful completion of any task would not be complete without the expression of appreciation of simple virtues to the people who made it possible.

I take this opportunity to express my deep & sincere gratitude to my project guide **Prof. Priyank Thakkar** , who have been kind enough to spare his valuable time on which we had no claim. His guidance & motivation conceived a direction in me, & made this project & the presentation successful.

A special thank is expressed wholeheartedly to **Dr K Kotecha**, Hon'ble Director, and **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department and **Prof. Gaurang Raval, PG ICT - Coordinator**, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation is extended throughout course of this work. And also for to provide basic infrastructure and healthy research environment.

I am also grateful to my Institution, all my faculty members and my colleagues specially Darshana and Parita in CSE Department for their special attention and suggestions towards the project work.

The blessings of God and parents make the way for completion of Project without whom I could not have made it here. I am very much grateful to them.

My deepest and heartly thank to my kids **(Amogh and Agrata)** and my spouse, without their support I could not complete this.

- Rinki Patel(11MICT53)

Abstract

In current times, it has gotten extremely vital for e-business organizations to enable their end clients to write opinion about the items/services that they have used. Such reviews provide vital sources of information on these products/services. This data is used by the future potential clients before choosing buy of new items or services. These opinions or reviews are also exploited by marketers to find out the drawbacks of their own products/services and alternatively to find the very important information related to their competitors products/services. Unfortunately this significant usefulness of opinions has also raised the problem for spam, which contains forged positive or spiteful negative opinions.

A recently proposed opinion spam detection method which is based on n-gram techniques is extended by means of feature selection and different representation of the opinions.The problem is modelled as the classification problem and nave-Bayes classifier and least-square support vector machine (LS-SVM) are used on three different representations (Boolean, bag-of-words and term frequency inverse document frequency (TF-IDF)) of the opinions. All the experiments are carried out on widely used gold-standard dataset and got encouraging result.

Also proposed a method to learn a classifier for the task of opinion spam detection in the presence of only small number of positive opinions (e.g. spam opinions). This method is inspired by a methodology named learning from positive and unlabelled examples. Results demonstrate that the proposed method gives good f-measure even in the presence of only small number of positive examples to learn a classifier.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In current times, individuals use web for all, they utilize web to settle their inquiries, to find results of unsolved issues, to think about not all that known items/services and so forth. They additionally utilize web, to know opinions of others before finalizing their choice on buying an item/service. A positive review about items for the most part brings about a buy of an item and the other way around. This reveals that opinions influence decision making of individuals and organizations. However, the significant influence of opinions in decision making has also encouraged spammer and is also the reason behind the increasing number of opinion spams.

Positive opinions can bring about significant financial additions distinction for business, organizations and people. Although negative conclusions on a few elements can harm their notorious. Deceptive opinions are deliberately composed to sound reliable and victimize readers. The task of deceptive opinion spam detection can be modelled as binary classification problem with two classes, deceptive and truthful.

The work focuses on modelling deceptive spam detection task as binary classification problem with deceptive and truthful as two classes. Impact of representation of the opinions in terms of different information retrieval models and feature selection is studied. Experiments are carried out on gold-standard dataset with nave-Bayes and LS-SVM classifiers.

1.1 Introduction of Opinion Mining

Opinion Mining is the field of study that analysis people's opinion. Opinion mining is the application of Natural language processing and text analysis to classify and mine subjective information in source materials.

Opinion mining mainly focus on opinions which express positive or negative sentiments. Opinion mining is subset of web mining, web mining is the application of data mining techniques to discover patterns from the web.

Web mining is the subset of data mining, the over all goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for future use.

Hierarchy of opinion mining is shown below.

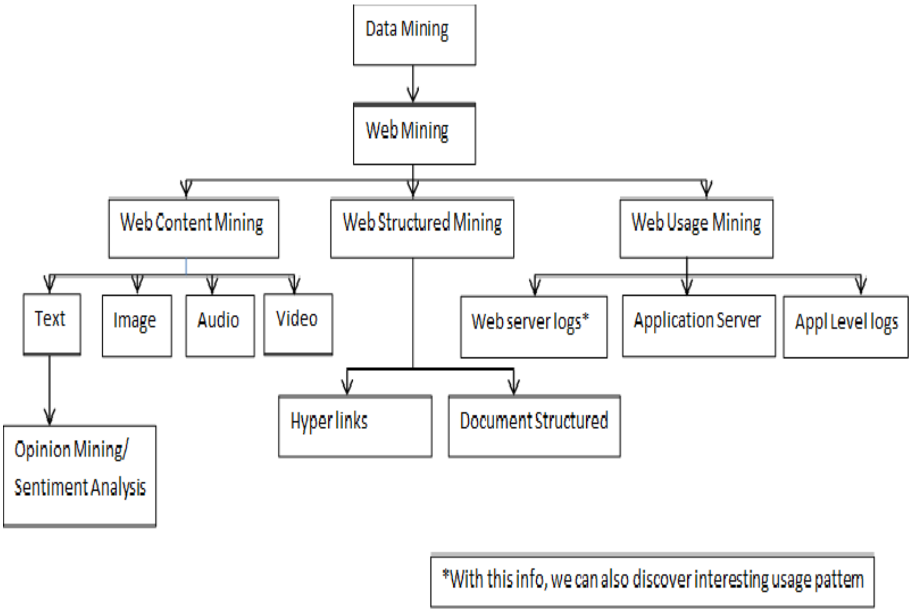


Figure 1.1: Opinion Mining Texonomy

1.2 Importance of Opinions

Opinions are always very much important for human being. Whenever we need to make a decision, we want to hear others opinions. In the past, to take individual decision we asked to our friends, relatives etc. And in business, they took surveys; focus on groups; and hired consultants.

Now a days people become technocrat, so uses of internet is increased and it is a common practice to buy product online and post their reviews on blog, forums and discussion groups.

There are two type of Opinions.

- **Positive opinions** Reviewer or company people wrote real or fake positive review about particular product/ service that can result in important financial grow and/or reputation for businesses, companies and individuals.
- **Negative opinions** Reviewer or company people purposely wrote negative review to some other product/ service in order to spoil their position.

There are two main objectives for writing spam reviews:

- To endorse some target product, people wrote unworthy positive review called hype spam.
- To spoil the position of some other target (e.g Competitors product), people wrote unfair negative reviews called defaming spam .

1.3 Types of Opinion Spam & Spammer [3]

There are three type of opinion spam

Type 1 (Untruthful(fake) reviews or product reviews)

People used positive review to promote one's own products called **Hype Spam**. Also used negetive review to defame ones competitors products called **Defaming Spam**.

Type 2 (reviews about brands only):

People wrote positive or negative reviews for company but not for particular product/ service. EX: I dont trust HP and never bought anything from them

Type 3 (non-reviews):

Non-reviews are those review which consider i) advertisements ii) other irrelevant tests containing no opinions

Types 2 and 3 spam reviews are easy to detect using supervised learning.

There are two type of Spammer

Individual Spammer: A person who writes fake reviews using his single user-id. e.g emplyee of company, person who write his own book.

Group spammers: Persons who works together to endorse a target entity and/or to spoil the position of another. The individual spammers in the group may or may not know each other.

A single person registers multiple user-ids and spam using these user-ids.

1.4 Charactristics of Opinion Spam Detection[5]

- **Review Centric Spam Detection:** Here spam detection is based only on reviews. A review has two main parts: rating and content.
  - **Compare content similarity:** To have the most extreme effect, a spammer may compose different surveys on the same item (utilizing different client ids). He/she might likewise compose reviews at numerous reviews destinations. Be that as it may, for a solitary spammer to compose different surveys that look altogether different is not a simple undertaking. Hence, a few spammers basically utilize the same review or slight varieties of the same review. In a late investigation of surveys from amazon.com, it was discovered that a few spammers were lazy to the point that they basically duplicated the same review and glued it for some different results of the same brand.

  - **Detect rating and content outliers:** If we reviews of a product contain only a very small amount of spam, then may we can detect possible spam activities based on rating behavior.

    In the event that an item has an extensive extent of spam reviews, it is difficult to distinguish them focused around survey evaluations, despite the fact that every spammer may act autonomously, on the grounds that they are no more outliers. This case is like gathering spam, which is additionally difficult to distinguish focused around substance alone on the grounds that the spam reviews are composed by distinctive parts of the gathering and there are an extensive number of them. Henceforth, their reviews are not anticipated that will be outliers.

- **Reviewer Centric Spam Detection:** Here,"unexpected" practices of analysts are misused for spam location. It is expected that all the reviews of every reviewer at a specific site are known. Most review destinations give such data, e.g., amazon.com, or such data might be found by matching client ids.

  - **Watch early reviewers:** Spammers may regularly the first few commentators to survey an item in light of the fact that prior reviews have a tendency to have a greater effect. Their assessments for every item are in one of the two ends, either good or bad. They may do this constantly for different results of the same brand.

  - **Detect early remedial actions:**For a given item, when somebody composes an (or the first) negative review to an item, the spammer gives a positive review simply after it, or the other way around.

  - **Compare review ratings of the same reviewer on products from different brands:**A spammer frequently composes exceptionally positive opinions for results of one brand (to push the item) and extremely negative opinions for comparable results of an alternate brand. A rating (or substance) examination will indicate inconsistencies. If some of the ratings also move away from the average ratings of the products, this is a good pointer of possible spam.

  - **Compare review times:** A spammer reviews a few products from different brands at nearly the same time. Such behaviors are irregular for a normal reviewer.

- **Server centric spam detection:** The server log at the review site could be useful in spam detection too. In the event that a solitary individual registers numerous times at a website having the same IP address, and the individual likewise composes various audits for the same item or even diverse items utilizing distinctive client ids, it is decently sure that the individual is a spammer. Utilizing the server log might additionally recognize some gathering spam exercises. For instance, if most great surveys of an item are from a specific area where the organization that prepares the item is found, it is a great sign that these are likely spam. As more individuals and associations are utilizing presumptions on the Web for choice making, spammers have more motivators to express false suppositions in item surveys, talks and online journals. To guarantee the nature of data gave by a presumption mining and/or inquiry framework, spam recognition is a basic assignment. Without compelling discovery, conclusions on the Web may get pointless.

1.5 Motivation

In todays world, it is growing trend to write online review about product/ service after utilizing it. People also read online review when they want to buy any product/ services, as it is very important for user as well as for companies too. And as we know there is no quality control on web, anyone can write anything in reviews, blogs etc., so we got low quality reviews and worse spam reviews. So to identify spam review is very essential task.

Also for human, it is very difficult to identify spam review. According to technology research firm garnter, there will be approx 10-15 percentage of online reviews will be fake or paid by 2014.

1.6 Thesis Organization

The thesis is organized as follows.
Chapter 1 (Introduction:) This chapter includes brief introduction about opinion and their type. Also mentioned characteristics of Opinion. The requirements of opinion spam detection is also mentioned.
Chapter 2 (Literature Survey:) Literature survey of different approaches for opinion spam detection are mentioned in this chapter. Also mentioned analysis of Opinion spam detection methods.
Chapter 3 (Opinion Spam Detection:) This chapter identified problem after literature survey and explained approach to solve Opinion spam detection task. And also explained about classification, IR model and n-gram technics which is used in work.
Chapter 4 (Experimental Evaluation:)This chapter explalined used tools, dataset, pre-processing and evaluations measures carried out during work. Also explained proposed algorithm for feature selection.
Chapter 5 (Learning from Positive and Unlabeled Example:) Chapter described PU-learning technique for opinion spam detection and Proposed PU-learning algorithm is also discussed.
Chapter 6 (Result and Discussion:) Results obtained after the implementation of proposed algorithm and analysis of the results is described in this chapter.
Chapter 7 (Conclusion and Future Scope:) This chapter concluding remarks of the work done and future scope for further optimization of the proposed algorithm is mentioned.

Chapter 2

Literature Survey

The various research papers related to Opinion Spam Detection was carried out during the dissertation phase. The basic highlights of some of the research papers and survey has been mentioned in this chapter. The details are as mentioned below.

- **"Opinion Spam and Analysis**[1]**"**

  In this paper, Supervised learning method was used for individual review using amazon.com database. They assumed duplicate and near duplicate reviews as spam opinion and other are non-spam. To find duplicate reviews, they considered i) review with same userid and same product, ii) Different userid and different products, iii) same user id and different products, iv) different userid and different products.

  In this paper, Shingle method (n-gram techniques) used to detect duplicate reviews. And Logistic Regression (AUC), Nave Bayes classification used to classify spam and non spam review.Still their is scope of to improve accuracy and detection method.

- **"Detecting Group Review Spam & Reviewer** [4]**,Spotting Fake Reviewer Groups in Consumer Reviews** [6] **"**

  In above papers, Supervised classification was used on amazon reviews and their reviewers. Here they used following criteria to find out group behavior.

i) Time Window, ii) Group Deviation, iii) Group Content Similarity, iv) Member Content Similarity, v) Early Time Frame, vi) Ratio of Group Size, vii) Group Size, viii) Support count
They used frequent pattern mining technique to find candidate groups, also used above criteria to find typical behaviors of groups and give rank to candidate groups using SVM rank, GS rank.

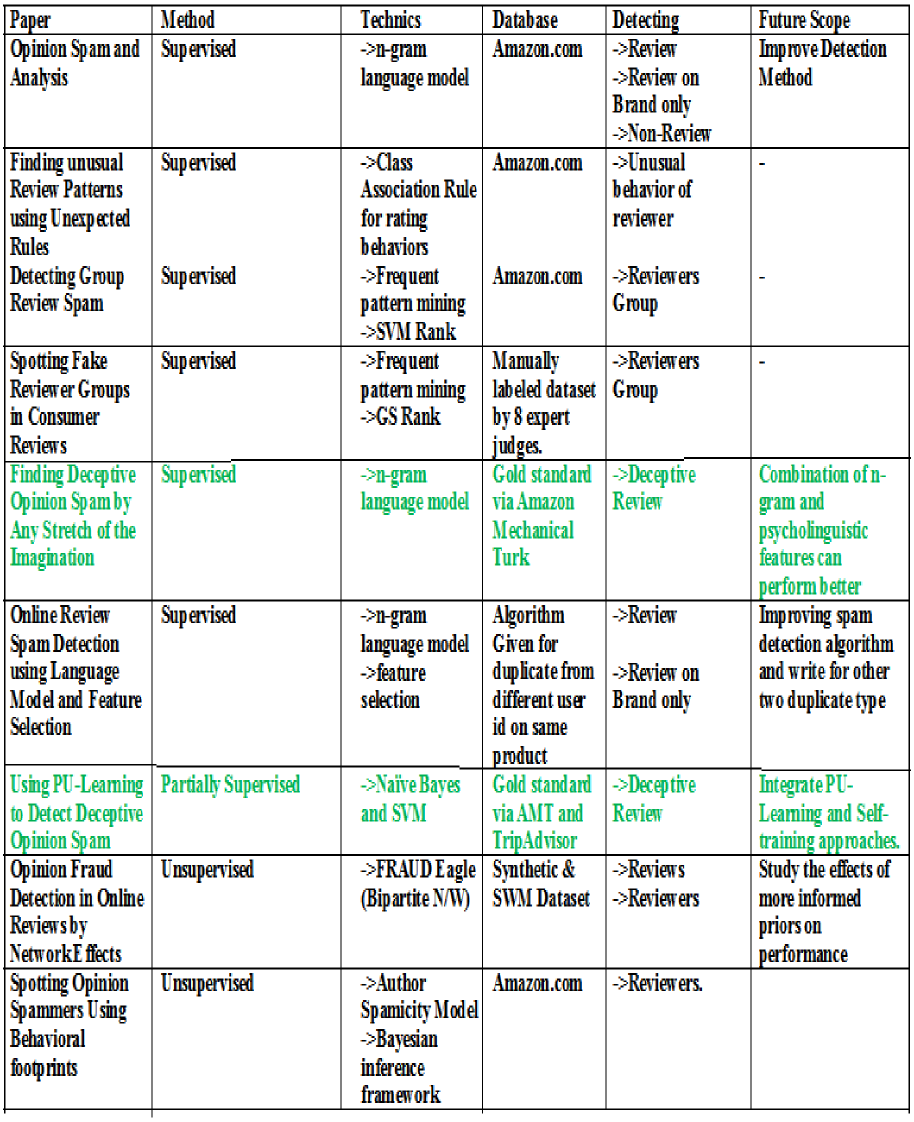
- **"Finding Deceptive Opinion Spam by Any Stretch of the Imagination[2]"**
  In this paper, Supervised learning method was used and first time Gold standard dataset consisting of 400 truthful and 400 deceptive hotel reviews was used which was created by Amazon Mechanical Turk (AMT).
  They used n-gram techniques for text categorization, LIWC (Linguistics Enquiry and Word Count) to detect psychological processes and part-of-speech was used to find out frequency distribution.
  In this paper, Nave Bayes and Support vector machine classifiers was used to train model using above techniques. To imporve spam detection, we can extended evaluation method.

- **"Using PU-Learning to Detect Deceptive Opinion Spam[7]"**
  In this paper, semi-supervised technique(PU-Learning) was used. And Nave Bayes and Support vector machine classifiers was used to train model uses a small set of deceptive opinion examples and a set of unlabeled opinions.
  To extended this proposed method, we can integrated the PU-learning and self-training approaches.

- **"Online Review Spam Detection using Language Model and Feature Selection[8]"**
  In this paper, supervised learning techniques using Language model and feature selection was used. First they collect review from the internet and did preprocessing. After that did POS tagging on reviews and finding duplicate

review given by different user id on same product to detect spam reviews. Also identified spam detection on other two types. And then joined result and do the analysis of review, is spam or not spam.

2.1 Analysis of Opinion Spam Detection Methods



Paper	Method	Technics	Database	Detecting	Future Scope
Opinion Spam and Analysis	Supervised	->n-gram language model	Amazon.com	->Review ->Review on Brand only ->Non-Review	Improve Detection Method
Finding unusual Review Patterns using Unexpected Rules	Supervised	->Class Association Rule for rating behaviors	Amazon.com	->Unusual behavior of reviewer	-
Detecting Group Review Spam	Supervised	->Frequent pattern mining ->SVM Rank	Amazon.com	->Reviewers Group	-
Spotting Fake Reviewer Groups in Consumer Reviews	Supervised	->Frequent pattern mining ->GS Rank	Manually labeled dataset by 8 expert judges.	->Reviewers Group	-
Finding Deceptive Opinion Spam by Any Stretch of the Imagination	Supervised	->n-gram language model	Gold standard via Amazon Mechanical Turk	->Deceptive Review	Combination of n-gram and psycholinguistic features can perform better
Online Review Spam Detection using Language Model and Feature Selection	Supervised	->n-gram language model ->feature selection	Algorithm Given for duplicate from different user id on same product	->Review ->Review on Brand only	Improving spam detection algorithm and write for other two duplicate type
Using PU-Learning to Detect Deceptive Opinion Spam	Partially Supervised	->Naïve Bayes and SVM	Gold standard via AMT and TripAdvisor	->Deceptive Review	Integrate PU-Learning and Self-training approaches.
Opinion Fraud Detection in Online Reviews by NetworkEffects	Unsupervised	->FRAUD Eagle (Bipartite N/W)	Synthetic & SWM Dataset	->Reviews ->Reviewers	Study the effects of more informed priors on performance
Spotting Opinion Spammers Using Behavioral footprints	Unsupervised	->Author Spamicity Model ->Bayesian inference framework	Amazon.com	->Reviewers.	

Figure 2.1: Analysis of Opinion Spam Detection Methods

Chapter 3

Opinion Spam Detection

Opinion spam detection is classified as binary classification problem which has two class Deceptive opinion and Truthful opinion. Deceptive spam opinion is those opinion which are written purposely to sound real and misguide reader. Deceptive spam opinion can be positive reviews or negative reviews.

Opinion spam detection has their own practical challenges which lead researcher to find more efficient method to find out fake review. As discussed in previous chapter, various approaches have been made towards suggesting fake review. But there is still scope of improvement in the performance of identify fake review.

3.1 Approach

We have modeled the opinion spam detection using supervised techniques named Naive Bayes and Least Square SVM. Also used n-gram techniques(unigram, Bigram and Bigram plus) with three representation of Information retrival model(Boolean, Bag-of-word, TFIDF). N-gram techniques is extended with feature selection approach. For feature selection, we used information gain.

We also used PU-learning technique to detect opinion spam. As we know, for text classification, system uses a set of labeled documents of n class as training set to build a classifier, which is then used to classify new documents into the n classes. For this we need labeled documents of all class in huge amount.

Whereas in PU-Learning, only small set of positive documents are required.

3.2 Classification Methods

In this work we used Naive Bayes and LS-SVM classifier to construct our model.

3.2.1 Naive-Bayes Classifier

Naive-Bayes classifier assumes class conditional independence. Given test data Bayesian classifier predicts the probability of data belonging to a particular class. To predict probability it uses concept of Bayes' theorem. Bayes' theorem is useful in that it provides a way of calculating the posterior probability, $P(C|X)$, from P(C), $P(X|C)$, and P(X). Bayes' theorem states that

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{3.1}$$

Here $P(C|X)$ is the posterior probability which tells us the probability of hypothesis C being true given that event X has occurred. In our case hypothesis C is the probability of belonging to class deceptive or truthful and event X is our test data. $P(X|C)$ is a conditional probability of occurrence of event X given hypothesis C is true. It can be estimated from the training data. The working of naive Bayesian classifier, or simple Bayesian classifier, is summarized as follows.
Assume that, 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 highest probability. By Bayes' theorem (Equation (4)),

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{3.2}$$

Given data sets with many attributes (A1, A2, ...,An), it would be extremely computationally expensive to compute $P(X|Ci)$. In order to reduce computation in evaluating $P(X|Ci)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e. that there are no dependence relationships among the attributes).Therefore,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{3.3}$$

$$= P(X_1|C_i) \times P(X_2|C_i) \times ... \times P(X_n|C_i) \tag{3.4}$$

Here Xk denotes to the value of attribute Ak for tuple X. Computation of $P(Xk|Ci)$ depends on whether it is categorical or continuous.

Distribution used for implementation is as follows:

- Multivariate Multinomial Distribution (MVMN): This distribution is used for Boolean representation matrix.
  For each feature model with a mvmn distribution, the Naive Bayes classifier computes a separate set of probabilities for the set of feature levels for each class.

- Multinomial Distribution (MN): This distribution is used for word count representation matrix.
  MN distribution is appropriate when all features represent counts of a set of words. The classifier counts the set of relative token probabilities separately for each class.

- Normal (Gaussian) Distribution: This distribution is used for tfidf representation matrix.

Normal distribution is appropriated for feature that have normal distributions in each class. The classfier estimates a separate normal distribution for each class by computing the mean and standard deviation of the training data in that class.

3.2.2 Least Square - SVM

A support vector machine (SVM) is another type of learning system, which has many desirable qualities that make it one of most popular algorithms. It not only has a solid theoretical foundation, but also performs classification more accurately than most other algorithms in many applications, especially those applications involving very high dimensional data. In general, SVM is a linear learning system that builds two-class classifiers.

To build a classifier, SVM finds a linear function of the form
f(x) = w:x+b so that an input vector xi is assigned to the positive class if f(xi) >= 0, and to the negative class otherwise. In essence, SVM finds a hyperplane w.x+b=0 that separates positive and negative training examples. This hyperplane is called the decision boundary or decision surface. Kernel functions are used for non-linear SVM.

Least squares support vector machines (LS-SVM) are least squares version of SVM, which are set of related supervised learning methods that analyze data and recognize patterns, and which are used for classification and regression analysis. In LS-SVM, we finds solution by solving a set of linear equations instead of a convex quadratic programming problem for classical SVMs. LS-SVM is a class of kernel-based learning methods.

3.3 Information Retrieval Model

- **Boolean:**Each element of the vector represents absence or presence of the corresponding term in the corresponding opinion.

  In Boolean model, each term is either present (1) or absent (0).

- **Bag-of-Word:**It represents a natural number indicating how many times the corresponding term has appeared in the corresponding review.

- **Term Frequency - Inverse Document Frequency (TFIDF):**

  TFIDF, is a numerical statistic that reflects how important a word is to a document in a collection. These vectors are real-valued vectors when represented in terms of TFIDF values of the terms. TF, IDF and TFIDF are defined in equations (3.1),(3.2) and (3.3) respectively [11].

  As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in.

$$TF = log(1 + f_{ij}) \tag{3.5}$$

$$IDF = f_{ij} * log(\frac{number of opinions}{number of opinions with word I}) \tag{3.6}$$

$$TFIDF = TF * IDF \tag{3.7}$$

Where, fij is the frequency of word i in opinion j.

3.4 N-Gram Techniques

N-gram is a coterminous grouping of n items from a given arrangement of content. The thing might be phonemes, letters, words, syllables or base sets as indicated by the provision.
Here n-grams typically collected from opinions.

- **Unigram:**An n-gram of size 1 is referred to as a "unigram".
- **Bigram:**An n-gram of size 2 is referred as "bigram"
- **Bigram Plus:**Combination of unigram & bigram called bigram plus.

Example: In this work we used unigram, bigram and bigram plus.

Unigram gives In, this, work, we, used, ... plus tokens.
Bigram gives In this, this work, work we, we used, ... bigram plus tokens.
Bigram plus gives In, In this, this, this work, work, work we, ... ,plus, bigram plus tokens.

Chapter 4

Experimental Evaluation

4.1 Tools

- **Weka 3.7**

  Weka (Waikato Environment for Knowledge Analysis) is a collection of machine Learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. **Usage:** Weka is especially used for pre-processing of the dataset.

- **Matlab 7.10 R2010a**

  MATLAB (matrix laboratory) is a high level fourth generation programming language and interactive environment for numerical computing. Developed by MathWorks, the language, tools, and built-in math functions of MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, and Fortran.

  Usage: Matlab is used for implementing supervised method e.g Naive Bayes and SVM-LS on the choosen dataset.

4.2 DataSet

Gold-standard English dataset assembled by the authors in [2] & [9] is used in this study. The dataset consist of 800 positive and 800 negative reviews. Out of 800 positive reviews, 400 reviews are truthful while remaining reviews are deceptive. Same is the case with negative opinions.

4.3 Pre-Processing

Dataset contents 1600 reviews and all reviews are in text file. Following process has been done to obtain input file and to do implementation.
All the characters are converted to lower case and stop words are removed from each of the reviews. Each of the review is then defined as vector in multidimensional Euclidean space.
Create different representation of each review using Wekas StringToWordVector tool. The axes of this multidimensional Euclidean space are terms appearing in opinion collection. Three different representations namely Boolean, bag-of-words and TFIDF of these vectors are exercised in this study.

In addition to this, sequence of words approaches, such as, unigram, bigram and bigram+ are used for each of the representations. The total number of features exhibited by unigram, bigram and bigram+ profiles of these opinions are 9378, 82093 and 92054 respectively.

4.4 Evalution Measures

The assessment of the usefulness of the proposed method was carried out by way of the f-measure. F - measure is defined as a harmonic mean of precision(P) and recall(R) values. Harmonic mean is more intuitive than the arithmetic mean when computing a mean of ratios. Computation of these evaluation measures requires estimating Precision and Recall which are evaluated from True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). These parameters are defined in Equations 3.5, 3.6, 3.7 and 3.8. And for these parameters is measure using following Confusion Matrix.

Confusion Matrix:

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

$$Precision_{positive} = \frac{TP}{TP + FP} \tag{4.1}$$

$$Precision_{negative} = \frac{TN}{TN + FN} \tag{4.2}$$

$$Recall_{positive} = \frac{TP}{TP + FN} \tag{4.3}$$

$$Recall_{negative} = \frac{TN}{TN + FP} \tag{4.4}$$

Precision is the weighted average of precision positive and negative while Recall is the weighted average of recall positive and negative. F-measure is estimated using equation 3.9.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.5}$$

4.5 Experimental Methodology

The task of detecting deceptive opinion is modelled as binary classification problem in this study. The two classes considered are deceptive and truthful. Two popular machine learning techniques namely nave-Bayes and LS-SVM are used to learn the classifier with 5 fold cross validation.

As reported in the previous section, each of the opinion is represented by large number of features. All the features may not be important and useful in distinguishing whether the opinion is truthful or deceptive. Information Gain is used to measure the importance of the features.

4.5.1 Proposed Algorithm using Feature Selection

As discussed, we want to classified opinions into Deceptive or Truthful opinions. For that first we do pre-processing on our data set. And convert data set in to three different representations called Boolean, Bag-of-words and TFIDF. Also construct Unigram, Bigram and Bigram Plus profile for each of the representation of each of the opinions. After that as we used Nave Bayes and SVL-LS supervised classifier to train model, so constuct training and testing set. Also applied feature selection on training set using information gain. And for that selected top N attributes based on information gain from training set. Where interval of attributes are 5, 50,100, 200, 500, 1000, 2000, 3000 up to N. N is total no of attributes which is 9375 for Unigram, 82093 for Bigram and 92054 for Bigram Plus. After that using the learned classifier predict for each of the opinion whether is Deceptive or Truthful.

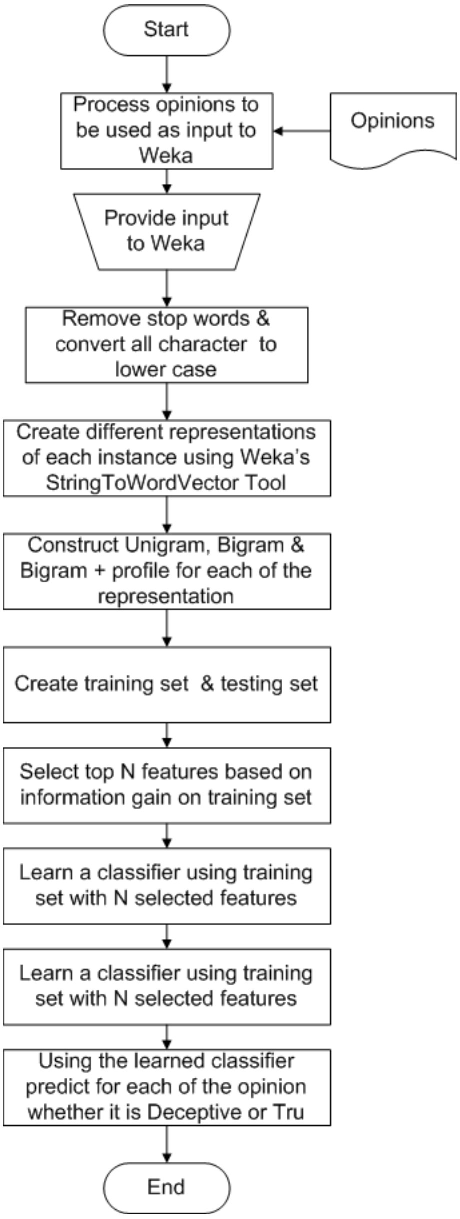

Start
Process opinions to be used as input to Weka
Opinions
Provide input to Weka
Remove stop words & convert all character to lower case
Create different representations of each instance using Weka's StringToWordVector Tool
Construct Unigram, Bigram & Bigram + profile for each of the representation
Create training set & testing set
Select top N features based on information gain on training set
Learn a classifier using training set with N selected features
Learn a classifier using training set with N selected features
Using the learned classifier predict for each of the opinion whether it is Deceptive or Tru
End

Figure 4.1: Flowchar of Proposed Algorithm

Chapter 5

Learning from Positive and Unlabeled Examples

5.1 PU-Learning

Text classification is an important problem that has many applications. As we know, for text classification, system uses a set of labeled documents of n class as training set to build a classifier, which is then used to classify new documents into n classes. For this we required huge labeled data consist of all class of data. If we talked about binary class then we required huge amount of labeled positive and negative data. But as in real scenario it is some time very difficult to find negative data. So we used PU-learning techniques where we required only small set of positive documents.

Key feature of PU learning method is no need of negative labeled document for learning.

Building Classifiers: Two-Step Approach [7]

- Identifying a set of reliable negative documents (denoted by RN) from the unlabeled set U.

- Building a classifier using P, RN and U-RN . This step may apply an existing learning algorithm once or iteratively depending on the quality and the size of the RN set.

5.2 Experimental Methodology

Gold-standard English dataset assembled by the authors in [2] & [9] is used in this study. The dataset consist of 800 positive and 800 negative reviews. Out of 800 positive reviews, 400 reviews are truthful while remaining reviews are deceptive. Same is the case with negative opinions.

In order to simulated real development to test our method we collected several different sub dataset from dataset. First we arbitrarily selected 160 deceptive opinions and 160 truthful opinions to build a fixed test set. The remaining 1280 opinions were used to build seven training sets of different sizes and distributions. They contain 20, 40, 80, 100, 150, 200 and 300 positive instances (deceptive opinions) respectively. In all cases we used a set of 980 unlabeled instances a distribution of 640 truthful opinions and 340 deceptive opinions.

5.3 Proposed Algorithm of PU-Learning

Figure 4.1 shows our adaption of the PU-learning approach for the task of opinion spam detection. In the proposed method first we find reliable negative instance from unlabeled data. First we took two matrices of N positive instance and unlabeled data. Then find mean of positive data set and after that find distance from unlabeled data instance to mean. And sort them in descending order and choose first N instance as reliable negative. Here N = 20, 40, 80, 100, 150, 200 , 300

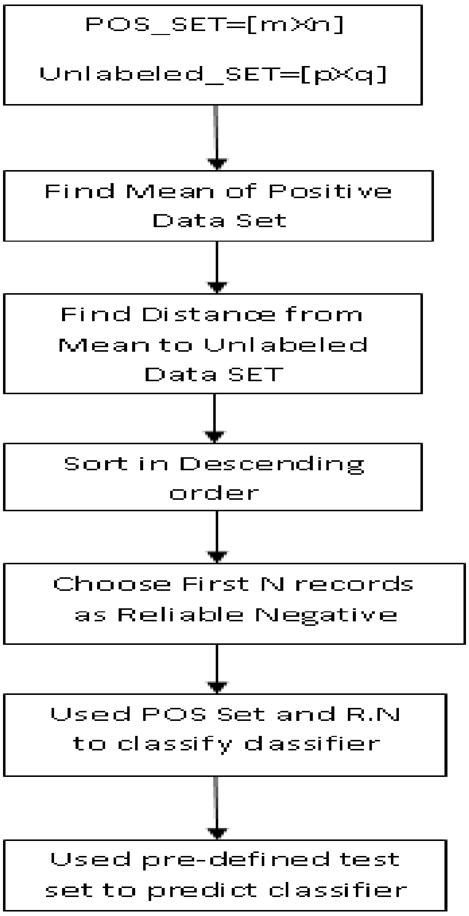

POS_SET=[mXn]
Unlabeled_SET=[pXq]
Find Mean of Positive Data Set
Find Distance from Mean to Unlabeled Data SET
Sort in Descending order
Choose First N records as Reliable Negative
Used POS Set and R.N to classify dassifier
Used pre-defined test set to predict classifier

Figure 5.1: Flowchart of Proposed Algorithm for PU-Learning

After finding reliable negative, we used Positive set and Reliable Negative data to classify classifier. Here we used Nave Bayes and SVM-LS as classifier. After that we used pre-defined test set to predict classifier.

Chapter 6

Result and Discussion

We have calculated precision, recall and f-measure as discussed in Chapter 4, to evaluate the framework of supervised techniques. Also evaluate proposed algorithm for feature selection and PU Learning.

6.1 Results of Supervised Techniques

Tables I, II and III show the result from all the experiments that were carried out. It is important to notify that Naive Bayes and SVM classifiers are used as learning algorithm. Also used Weka for pre processing of data and Matlab for implementation of mentioned supervised method.

Experimens is carried with n-gram(UniGram, Bigram, BigramPlus) on all three IR models (Boolean, Word count, TF-IDF).
Following is result of Two Class ('d', 't') d: Deceptive Opinion t: Truthful Opinion with Stratified (5 Fold) **Unigram Data, Attribute: 9378**

Table I show Highest F-measure in Nave Bayes on word count: **85.95**

	Boolean			Word Count			TFIDF		
	P	R	F	P	R	F	P	R	F
Naive Bayes	86.34	85.12	85.73	86.03	85.88	85.95	74.81	71.98	73.36
LS-SVM	79.11	78.81	78.96	78.57	78.25	78.41	79.60	79.13	79.36

Table I: Unigram Data

	Boolean			Word Count			TFIDF		
	P	R	F	P	R	F	P	R	F
Naive Bayes	82.03	74.88	78.29	85.18	82.38	83.75	76.69	76.05	76.37
LS-SVM	82.72	78.75	80.68	83.49	79.63	81.51	82.87	79	80.89

Table II: Bi-Gram Data

Table II show Highest F-measure in Nave Bayes on word count: **83.75**

	Boolean			Word Count			TFIDF		
	P	R	F	P	R	F	P	R	F
Naive Bayes	83.75	78.88	81.24	88.85	87.75	88.29	78.71	78.45	78.58
LS-SVM	85.72	83.31	84.5	84.79	82.31	83.53	84.6	82.12	83.34

Table III: BiGram Plus Data

From above Table-III we show that our reimplementation of Ott et al's model with Bigram plus obtained almost 89 % f-measure with word count using Naive Bayes classifier.

6.2 Result of Feature Selection

Use of n-gram techniques with feature selection using information gain was also made. Graphical results show that supervised text classification techniques using n-gram with feature selection (Information Gain) almost gave same f-measure with consideration of only few attributes.

The information gain is used for feature selection. And applied on all three representation of IR model.

Naive Bayes & SVM Classification is used on interval of attributes
e.g 5,50,100,200,500,1000,2000,3000...... up to N
N=9376 for UniGram
N=82093 for Bigram
N=92054 for BigramPlus

Following graphs are representation of F-measure using Nave Bayes & SVM-LS classifier.

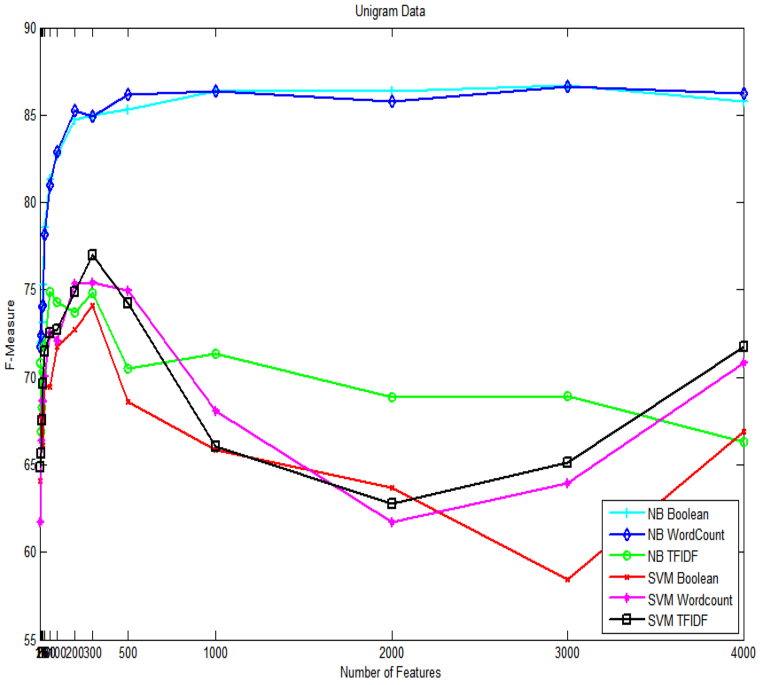


Figure 6.1: Unigram with no. of different features

From Fig 1, it was identified that considering only 1000 attributes also got f-measure 86.36 which is almost same to the original distribution obtained using all attributes.

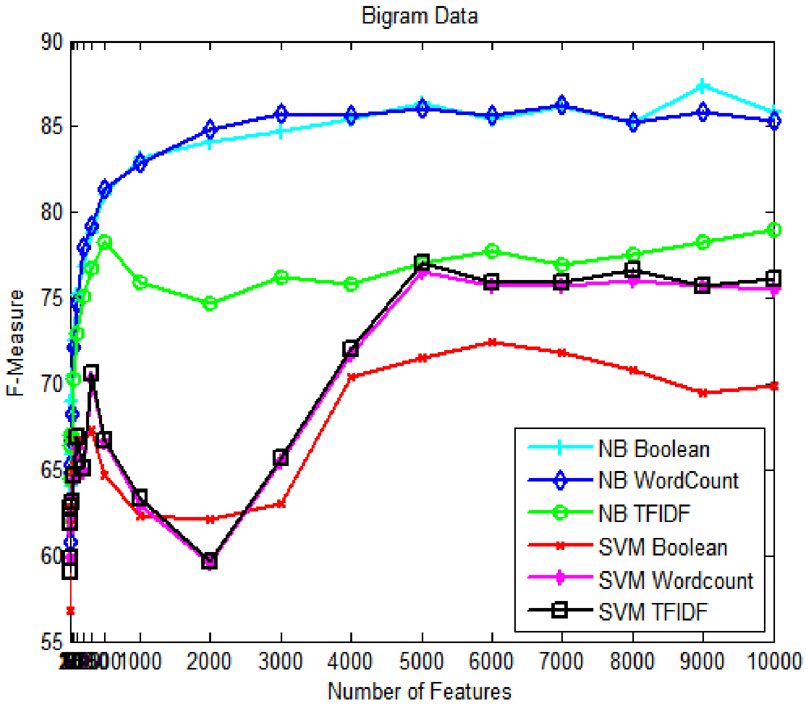


Figure 6.2: Bigram with no. of different features

Whereas Fig-3 shows almost same f-measure 89.40 at 9000 attributes.

Graphical results show that considering few attributes to build classifier with feature selection gave same result as considering all attributes in supervised techniques. We also conclude that after certain point if number of attributes increases that will not effect in increasing F-measure.

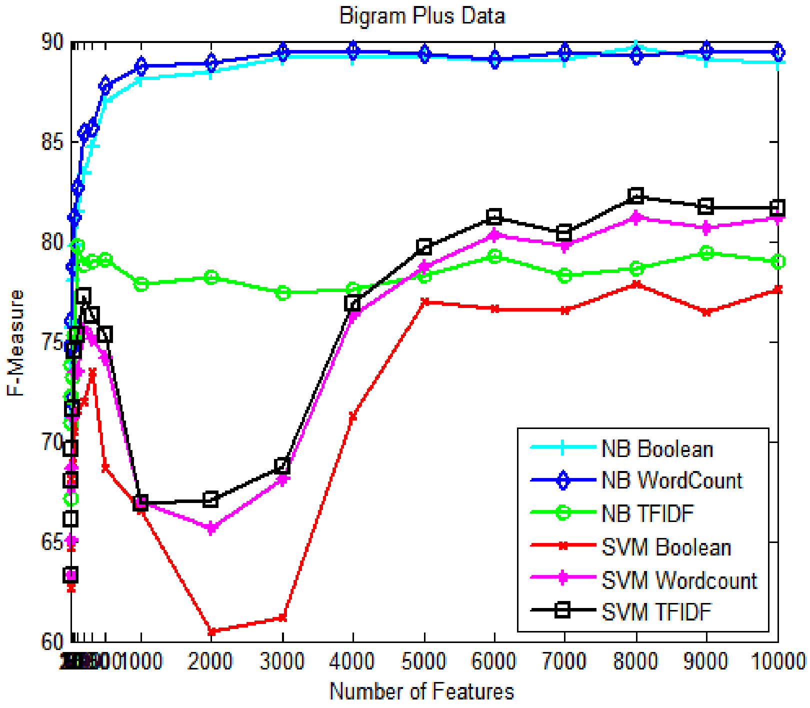


Figure 6.3: Bigram Plus with no. of different features

6.3 Result of PU Learning

Here we used Boolean and word count representation on Unigram data only.

Table show the results from all the experiments we carried out. It is important to notice that we used Naive Bayes and SVM classifiers as learning algorithm in our PU-Learning method.

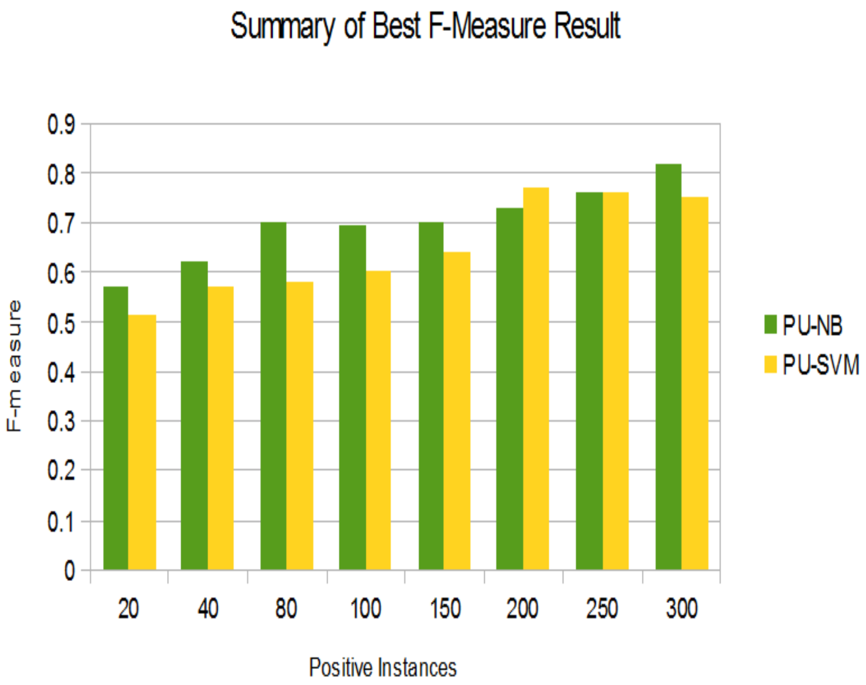


Figure 6.4: Chart of PU-Learning

Figure presents a summary of the best results obtained by each of the methods in dataset. It is important to notice that best result obtained by the proposed method is 0.82 f-measure in the deceptive opinion spam which is comparable best result (0.89) reported for this dataset. Here we only consider 300 positive instances which is almost 20% of all 1600 instances.

No. of Pos Data	**Approach**		**DECEPTIVE**			**TRUTHFUL**		
			P	R	F	P	R	F
20	NB	BL	0.50	0.68	0.57	0.53	0.97	0.68
		WC	0.51	0.99	0.67	0.82	0.05	0.09
	SVM	BL	0.67	0.41	0.51	0.57	0.79	0.67
		WC	0.72	0.38	0.50	0.58	0.86	0.69
40	NB	BL	0.60	0.63	0.62	0.61	0.58	0.59
		WC	0.53	0.95	0.68	0.76	0.16	0.26
	SVM	BL	0.62	0.53	0.57	0.59	0.67	0.63
		WC	0.63	0.44	0.52	0.57	0.74	0.65
80	NB	BL	0.57	0.91	0.70	0.76	0.31	0.44
		WC	0.57	0.91	0.70	0.76	0.31	0.44
	SVM	BL	0.67	0.51	0.58	0.60	0.75	0.67
		WC	0.64	0.47	0.54	0.58	0.74	0.65
100	NB	BL	0.58	0.87	0.69	0.74	0.37	0.49
		WC	0.58	0.87	0.69	0.74	0.37	0.49
	SVM	BL	0.70	0.52	0.60	0.62	0.78	0.69
		WC	0.66	0.48	0.56	0.59	0.75	0.66
150	NB	BL	0.59	0.87	0.70	0.76	0.39	0.52
		WC	0.59	0.87	0.70	0.76	0.39	0.52
	SVM	BL	0.70	0.58	0.64	0.64	0.75	0.69
		WC	0.69	0.53	0.60	0.62	0.76	0.68
200	NB	BL	0.78	0.59	0.68	0.67	0.83	0.74
		WC	0.63	0.97	0.76	0.93	0.43	0.59
	SVM	BL	0.71	0.84	0.77	0.80	0.65	0.72
		WC	0.71	0.76	0.73	0.74	0.68	0.71
300	NB	BL	0.76	0.89	**0.82**	0.87	0.72	0.78
		WC	0.67	0.97	0.79	0.95	0.52	0.67
	SVM	BL	0.69	0.82	0.75	0.78	0.64	0.70
		WC	0.69	0.78	0.73	0.75	0.66	0.70

Table IV: Result of PU-Learning

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

The task of opinion spam detection is focused in this work. The problem of opinion spam detection is modelled as the classification problem. Experiments are carried out with unigram, bigram and bigram+ representations of the opinions. For each of these representations, opinions are modelled as Boolean, bag-of-words and TFIDF vectors. Nave-Bayes and LS-SVM are used as the classification techniques. It is evident from the results that nave-Bayes classifier with bag-of words representation of the opinions performs the best. Impact of feature selection is also studied in the work. It is apparent form the result that learning a classifier using appropriate number of features improves the accuracy.

The work also proposes a method to learn a classifier for the task of opinion spam detection in the presence of only small number of positive opinions (e.g. spam opinions). This method is inspired by a methodology named learning from positive and unlabelled examples. Results demonstrate that the proposed method gives good f-measure even in the presence of only small number of positive examples to learn a classifier. This f-measure is comparable to the situation where we have large no of labeled examples.

7.2 Future Scope

- The Use of more than certain percentage of adjectives, adverb, and missing certain key facts about products or services can be applied as feature to achieve more accuracy in detection of opinion spam. Also plan is to do improvement in detection of spam review using POS tagging for adjective, noun and verbs etc.

References

[1] Nitin Jindal, Bing Liu, Opinion Spam and Analysis,ACM Proceedings of the international conference on Web search and web data mining, pp.219-229, 2008.

[2] MyleOtt, Yejin Choi, Claire Cardie, Jeffrey T. Hancock, Finding deceptive opinion spam by any stretch of imagination , ACM Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 , pp.309-319,2011.

[3] Nitin Jindal, Bing Liu ,Review Spam Detection , ACM Proceedings of the 16th international conference on World Wide Web, pp-1189-1190,2007.

[4] Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, Nitin Jindal, ”Detecing group review spam”, ACM Proceedings of the 20th international conference companion on World Wide web, pp.93-94, 2011.

[5] Bing Liu, Web Data Mining

[6] Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, ”Spotting fake reviewer group in consumer reviews”, ACM Proceedings of the 21st international conference companion on World Wide web, pp.191-200, 2012.

[7] Rafael Gazman Cabrera, Manuel Montes, Paolo Rosso, ”Using PU-Learning to Detect Deceptive opinion Spam”, Atlanta, Georgia, 14 June 2013. cO2013 Association for Computational Linguistics.

[8] Manali S Patil and A M Bagade. Article: Online Review Spam Detection using Language Model and Feature Selection. International Journal of Computer applications 59(7):33-36, December 2012.

[9] Ott, Myle, Claire Cardie, and Jeffrey T. Hancock. ”Negative deceptive opinion spam.” Proceedings of NAACL-HLT. 2013.

[10] Hall, Mark and Frank, Eibe and Holmes, Geoffrey and Pfahringer, Bernhard and Reutemann, Peter and Witten, Ian H, The WEKA data mining software: an update, ACM SIGKDD Exploration Newsletter, 11(1), 10-18, 2009.

[11] B.Azhagusundari, Antony SelvadossThanamani Feature Selection based on Information Gain, ISSN: 2278-3075,Volume-2, Issue-2, January 2013. In IJITEE

Index