



University of
Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: Master of Science in Computer Science	Spring semester, 2014..... Open / Restricted access
Writer: Samuel Daniel (Writer's signature)
Faculty supervisor: Erdal Cayirci External supervisor(s):	
Thesis title: Predictive modeling of trust to Social Media content	
Credits (ECTS): 30	
Key words:	Pages:98..... + enclosure: CD..... Stavanger, 27/06/2014..... Date/year

Predictive modeling of trust to social media content

Samuel Daniel

Faculty of Science and Technology
University of Stavanger

June 2014

Abstract

In recent years, social networking sites have got a massive popularity because they let people to devise a public profile within a tied system. As the popularity increases and they became widely used as one of the important sources of news, people become more cautious about determining the trustworthiness of the information which is disseminating through social media for various reasons. For this reason, knowing the factors that influence the trust in social media content became very important. In this thesis, we use a survey as a mechanism to study trust in social networks. First, we prepared a questionnaire which focuses on measuring the ways in which social network users determine whether content is true or not. And then we analyzed the response of individuals who participated in the survey and discuss the results in a focus group session. Then, the responses, we get from the survey and the focus group was used as a dataset for modeling trust, which incorporates factors that alter trust determination. The dataset had initially 108 records, but subsequent to preprocessing a total of 106 records were used for building the models.

In this study, linear regression, logistic regression, Poisson regression and negative binomial regression were applied on our dataset. According to the results of the various types of tests done on these models, we concluded that the logistic regression model is the most reasonably accurate regression model for trust in social networks. R and Minitab were the tools that were used for the analysis.

In this thesis, an endeavor was made to apply the Decision Tree, Bayesian Classifiers and Neural Network predictive data mining techniques in significant social media factors for predicting trust. To accomplish this goal: The WEKA data mining tool was used to evaluate the J48, Naïve Bayes and Multilayer Perception algorithms.

Distinct experiments were made by performing adjustments of the attributes and using various numbers of attributes in order to come up with a purposeful output. After comparing the resulting models using WEKA's experimenter we concluded that Multilayer Perception algorithms were the best suited classification model in comparison with Naïve Bayes and J48 algorithms.

Moreover, the most determinant factors when it comes to predicting trust were identified. Namely, these are Age, Years of use, Important news source, Favorite social network site, Gender and Number of people sharing. Overall, this research has verified that regression and data mining techniques are worthwhile to scale up the efficiency of trust modeling and prediction process.

Acknowledgement

I owe the deepest gratitude to Professor Erdal Cayirci, my advisor, who from the very beginning of the idea to its realization has given me his substantive guidance and feedbacks. His encouragement, passion, tolerance, unlimited support and giving valuable feedbacks are really appreciated.

I would like to thank all people who participated in the survey and focus group sessions. I also extend my gratitude to the staff members of Computer Science Department, UIS, for their unconditional assistance.

My deepest thanks goes to my family, who have been on my side all the way until I fulfill my dream, and I am also highly indebted to my friends for their everlasting support and encouragement throughout my study.

Table of Contents

ABSTRACT.....	iii
ACKNOWLEDGMENT.....	IV
CHAPTER ONE.....	2
1.1 INTRODUCTION.....	2
1.2 Research Problems.....	2
1.3 Objectives of the research	3
1.3.1 Specific Objectives.....	3
1.4 Organization of the research.....	3
1.5 Methodology.....	4
1.5.1 Questionnaire.....	4
1.5.2 Focus Group	4
1.5.3 Cluster modeling.....	5
1.5.4 Classifications.....	5
1.5.5 Regression analysis.....	5
CHAPTER TWO.....	6
2.1 Background.....	6
2.2. Data mining technique	7
2.2.1 Classification.....	7
2.2.2 J48 Decision Tree Algorithm.....	8
2.2.3 Neural Networks.....	8
2.2.4 Naïve Bayes	9
CHAPTER THREE.....	10
3.1 Source of data.....	10
3.2 Result of survey.....	11
3.3 Focus Group result.....	20
3.4 Selected attributes.....	22
CHAPTER FOUR.....	24
EXPERIMENTATION.....	24
4.1 Experiment Design.....	24
4.2 Cluster Modeling.....	24
4.2.1 Experimentation I.....	25
4.2.2 Experimentation II.....	27
4.2.3 Experimentation III.....	28
4.2.4 Experimentation IV.....	31
4.2.5 Selecting the best Clustering Model.....	32
4.3 Classification Modeling.....	33
4.3.1 J48 Decision Tree Model Building.....	34
4.3.2 Naïve Bayes Model Building.....	35
4.3.3 Neural Network.....	37
4.3.4 Chosen rules	38
4.3.5 Choosing the best Clustering Model.....	39
4.4 Regression Modeling.....	41
4.4.1 Linear Regression.....	41

4.4.2 Logistic Regression in Minitab.....	44
4.4.3 Logistic Regression in R.....	46
4.4 .4 Poisson Regression.....	48
4.4.5 Negative binomial Regression.....	52
4.5 Comparison of the regression models.....	56
 CHAPTER FIVE.....	 58
5.1 Conclusion and Future Works.....	58
 REFERENCES.....	 60
 APPENDIXES.....	 62

1. CHAPTER ONE

1.1 Introduction

In social networks people keep in touch with their friends by posting some kind of content in their walls and sharing news, clips and any kinds of activities they have inclination to and preserve their involvement on the social media. Forming new relationship in these sites doesn't have any limitation of both place and time, which makes it quite easy and attractive. This days the number of people who use social media as a source of news is increasing rapidly even though they have still to a certain extent a doubt about truthfulness of the contents which are propagated across the social network in a daily basis.

Since social networks are organized around the people who use them, trusting the content which is propagated in them is solely dependent on the determination ability of the users. If the users don't trust the information then he/she will not propagate it.

The main objective of this study is to assess the different ways of trust determination factors and to find the most important factors which can be used to model trust in social media content.

1.2 Research Problem

Even though the number of people who use social media as their most important news source is rising, the trust they have to social media content is comparatively low.

How can people successfully determine a trustworthiness of content in social media?

The main reason behind this problem is that until now there doesn't exist a mechanism to determine the trustworthiness of a content based on certain criteria. While doing the research certain topics become especially relevant in relation to the question above. The two questions listed below are some of the most relevant question with the problem stated above.

How much trust do you have in social media as a source of news? In a scale of 0 to 5 (5 if you fully trust them and 0 if you don't trust them at all).

Which of the following do you need to trust to social media content? (You can select multiple) Please also order these criteria from the most important to the least.

- The source is known and well reputed by you
- High number times the content is liked, shared and forwarded
- Verified by conventional media
- Verified by friends and colleagues
- Common sense or your intuition

The main focus of the study was on these topics, and the demonstration of the findings is therefore structured around these topics.

1.3 -The objective of the research

The main objective of this research is to design a predictive model for trust in social media networks by using regression and data mining techniques from the survey dataset that is capable of elevating the probability of determining trust to social media content.

1.3.1 Specific Objectives

- To distinguish and choose parameters or attributes which are highly significant with regard to trust modeling and prediction from the data set.
- To compare Linear, Logistic, Poisson and Negative Binomial regression methods to find the best regression model that fits our data set.
- To compare outputs of J48 Decision Tree, Bayesian Classifiers and Neural Network in order to find the best classification model to predict trust to social media content.
- To evaluate results of K-Means cluster algorithm by changing the values of the parameters to find the most efficient cluster model
- To explain and analyze the outputs of the chosen model.

1.4 - Organization of the Thesis

This study is organized into five chapters. The first chapter gives general overview of the problem area and the data mining technologies. It also describes the main and specific objectives of the thesis, limitations of the study and the importance of the results of this study.

The first chapter briefly discusses background to the problem area and DM technology, and states the problem, objective of the study, research methodology, scope and limitation, and significance of the results of the research.

Chapter two explains related literature reviews based on previous research done in the topic area.

The third chapter deals about the different data mining techniques and regression methods that were used in this study. It gives a brief explanation of decision tree, Naïve Bayes, Neural Networks, K-Means clustering algorithm and regression analysis methods.

In the fourth chapter a brief description of the experimentation results and analysis of the findings of the study were made. The clustering, classification and regression experimentation phases were included. Moreover, evaluation of the findings is also done.

The fifth and last chapter is allocated to concluding remarks and future plans for the study.

1.5 Methodology

This study uses two kinds of empirical methods, namely questionnaire and focus group to collect information. The questionnaire was chosen to collect information because it makes it easier to distribute to as many people as you want, but it is quite difficult to get a detailed analysis by using just the data which is collected by questionnaire. As a result, we decided to use the focus group method to supplement the information we get from the questionnaire by discussing with people who have information technology educational background and technical know-how of the research area.

After the data was collected, it was preprocessed and prepared in a way suitable for the data mining tasks. Then experiments were carried out in three sub phases, first the cluster modeling which was then followed by classification modeling and finally regression modeling phase.

In this study, WEKA (Waikato Environment for Knowledge Analysis) tool was used for clustering and classification purposes and, R and Minitab were used for Regression tasks.

1.5.1 Questionnaire

Before starting to write the questions which were used in the questionnaire we made extensive research by reading articles related to the topic of our project and in particular about “ trust “. After that we prepared the questions with the collaboration of the supervisor and sent out a hard copy version of the questionnaire for ten students to get a feedback mainly about the type of questions we used and their opinion about it.

Based on the feedback we get from them we reduced the number of questions in the questionnaire which was originally 27 to 23, and we also made changes on the ways of some of the questions were presented. Then we translated the questionnaire to Norwegian language to give people an option to use one of the two languages which they are comfortable with.

After that we created an account and sent out the questionnaire via surveymonkey.com, and distributed the link by using email and face book. The whole questionnaire can be seen in the appendix section. After the data was gathered, the diagrams were created and analyzed with the help of surveymonkey.com.

1.5.2 Focus Group

We arranged a group of 4 programmers who were former students of this particular masters program to be part of the session. This method was chosen because it makes it possible to get an impulsive response from participants at that particular time and to put into perspective whole different ideas that can be proposed by discussing the matter as a group.

At the beginning of the discussion I presented to the group the findings of the questionnaire, in order to give them an overview of the key findings. The discussion continued by raising some of the surprising findings of the questionnaire and the possible factors which made the participants to select them. The focus-group discussion was held in UIS.

1.5.3 Clustering

Clustering is a process of classifying a diverse collection of unlabeled data into several groups according to certain features in a data set.

The k-means clustering algorithm was used in this study, since it is easy to understand and to a good extent scalable, in addition its simplicity for transformation in order to deal with streaming data makes it a good choice. Even though, it's prerequisite of that the number of clusters should be specified before the algorithm is applied works against it.

1.5.4 Classification

As one of the main goals of this study is to predict trust using data mining techniques, a classification technique was adopted to develop a predictive model. The models were built with three different supervised machine learning algorithms i.e. Decision Tree Classification Algorithm, Bayesian Classifier and Neural Network using WEKA 3.6.11 machine learning software.

1.5.5 Regression Analysis

Regression analysis is one of the most often used tools in predictive modeling. It allows people to analyze the relationships between dependent and independent parameters. The dependent parameter is the one we really care about, whereas the independent parameters are the contributors for achieving those results.

In this study, four different kinds of regression analysis were made, and they were compared for their goodness of fits on the basis of AIC, log-likelihood and the two deviances (null and residual). Linear, Logistic, Poisson and Negative Binomial regression analysis were the type of regression analysis's that were used in this study.

CHAPTER TWO

2.1 Background

According to definition.net [20] trust means reliance on the integrity, strength, ability, surety, etc., of a person or thing; confidence.

“Trust is both an emotional and logical act. Emotionally, it is where you expose your vulnerabilities to people, but believing they will not take advantage of your openness. Logically, it is where you have assessed the probabilities of gain and loss, calculating expected utility based on hard performance data, and concluded that the person in question will behave in a predictable manner. In practice, trust is a bit of both. I trust you because I have experienced your trustworthiness and because I have faith in human nature.” [19]

As it is clearly described in the last sentence of the previous paragraph, most people trust others because they had experienced trustworthiness from them in their earlier interaction. So we can use this factor for modeling of trust in this study, because in social network sites this factor has a huge influence on trusting a content which is shared by people who have already got a credibility because of their previous posts quality. In social network sites the most important factors for building trust are reputation and influence. When we say reputation in social media it means the way you are perceived by others solely based on your posts. And your influence can be explained as the number of people who will reply or like the post in your wall.

According to Fogg (2000), having a trust indicates a positive belief in another person, or content in this particular case. Ordinary users are more likely to trust people who share information which is solely based on actual facts, like by attaching the links related to the contents they share, which will most likely increase the credibility of the information they share. Even though it is quite a new area of research there are some useful researches which are done in the last few years. Such as “Propagation Models for Trust and Distrust in Social Networks” by Cai-Nicolas Ziegler and Georg Lausen [21], proposes a model for both trust and distrust in social networks.

And also the researches made by likes of “Models and Methods in Social Network Analysis” by Carrington P. J., Scott J., and Wasserman S. (2005) [4] and “A Flexible Trust Model for Distributed Service Infrastructures” by Liu Y., Yau S., Peng D., and Yin Y. (2008)[6] were really helpful in introducing some of the already existing trust metrics.

2.2 Data mining techniques

In this chapter the detailed explanations of all the methods which are used for this research and the theories behind the various models of the analysis are described. This part also addresses the feasible probability distributions of trust data and their appropriate regression models. In addition, It also accommodate the explanation of the software packages which were used for this analysis and modeling.

2.2.1 Classification

This research uses classification techniques for predicting trust. The three types of classification techniques that were used to construct prediction models are Decision Tree(j48), Neural Network(Multilayer perception) and Bayesian(Naïve Bayes) Classifiers. Moreover, the three algorithms that were used to construct the models and the output matrices of the algorithms that were used to measure the performance of the algorithms and comparison are explained thoroughly.

As Han & Kamber (2006) have stated, classification have two distinct processes, namely learning and classification. Throughout the learning process, a classifier will be built portraying a set of beforehand determined classes that will later portrayed in the form of classification rules. The classification algorithm builds the classifier by analyzing a training set and their associated class labels.

2.2.2 Decision Trees

A decision tree is a data mining technique that generates a graphical illustration and analysis of the model it generates. The model that is generated by decision tree could be either predictive or descriptive model.

According to Alberto(2000,) even though decision trees are widely used for classification purpose , they can be used also for different kinds of regression analysis. Basically, building decision tree classifiers does not need in detail know how of domain knowledge or attribute setting, hence, it becomes quite popular for exploratory knowledge discovery. Decision trees can handle high dimensional data.

The illustration of the acquired knowledge in the form of a tree is quite straight forward to assimilate by anyone. As a result, the two steps of classification techniques in decision tree (learning and classification) are plain and quick, and also they have pretty good accuracy. Although, the type of data we have also has a huge importance when it comes to determine how successful it's in our usage.

To mention some of the application areas where Decision tree algorithms usage has been common are Pharmacology, Remote sensing, Software development, Physics, Agriculture and Medicine.

2.2.3 J48 Classifier Algorithm

J48 is an implementation of the well known C4.5 algorithm for producing either pruned or unpruned C4.5 tree. The C4.5 algorithm was built based on the concept of information obtaining or entropy reduction to select the most efficient split.

In general, It assumes that individual attributes of the data can be used to make a decision by splitting the original data into minor subsets.

The J48 decision tree algorithm is the one that is used in this research to classify the social media content as trusted or non-trusted.

The main reason J48 decision tree was chosen to serve as a model for classification is that it produces simpler rules and remove insignificant parameters before it begins a process of tree induction. Usually, J48 decision trees happen to had a relatively higher accuracy than other classification algorithms, In addition, J48 also provides extremely fast and pretty powerful way of fast and powerful way to show structures for a data.

2.2.4 Neural Networks

According to Quinlan (1993) Neural network make use of a multilayered approach which estimates sophisticated mathematical functions to process a specific data.

Neural networks are well known for their learning efficiency. They perform much better in comparison with the other classifier algorithms when the majority of variables are weakly relevant. One disadvantage of neural networks is that they took longer time to learn.

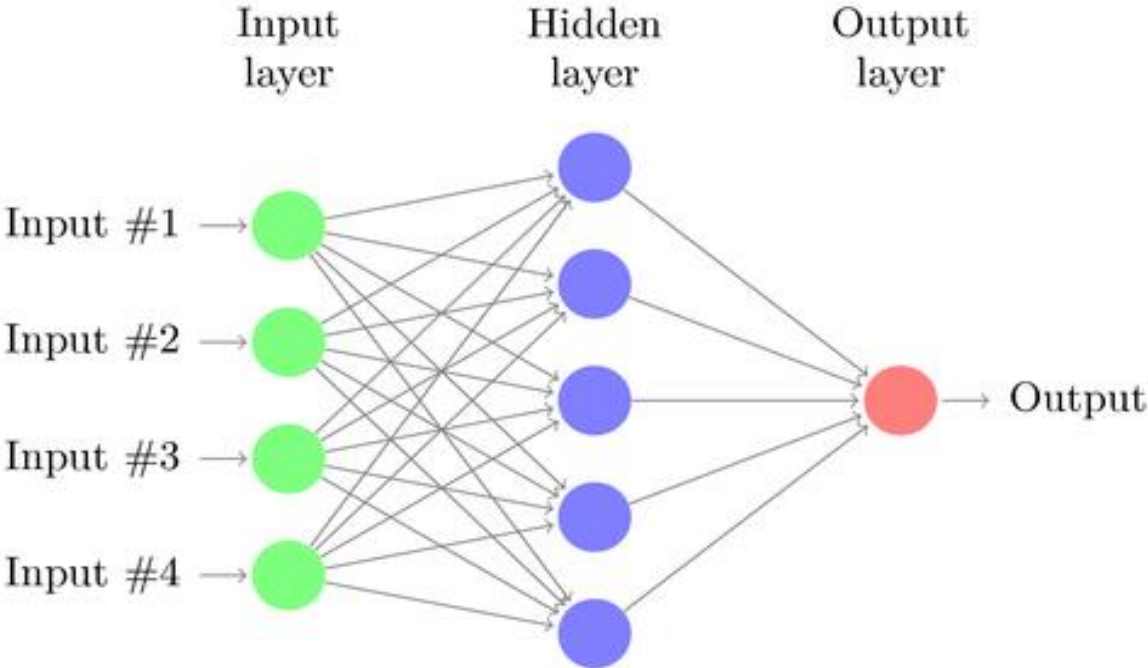


Figure 1 – Neural Network []

2.2.5 Naive Bayes

According to (Bhargavi & Jyothi, 2009) a Naïve Bayes classifier works under the assumption of that the presence of a specific feature of a class have no association to the presence of any other constituent.

The Naïve Bayes algorithm makes use of Bayes' Theorem, which is a formula that determines a probability by estimating the frequency of values and mixture of values in the previously collected data. It determines the probability of an event happening provided that the probability of another event that has already happened.

The Bayes' theorem is stated as follows

$$P(H/X) = P(X/H) P(H) / P(X)$$

The Naive Bayes algorithm provides a way to mix the prior probability and conditional probabilities within a single formula that can be used to determine the probability of each of the classifications in turn. After that, the class with the highest value will be chosen as the class of the new instance (39).

CHAPTER THREE

3.1 Source of data

The source of data for this research is my own data set, which is obtained by using a questionnaire and focus group to collect information. The questionnaire was chosen to collect information because it is easier to distribute to as many people as you want, however, it is quite difficult to get a detailed analysis by using just the data which is collected by questionnaire. As a result, we decided to use the focus group method to supplement the information we get from the questionnaire by discussing with people who have information technology educational background and pretty good technical know-how of the research area. Before starting to write the questions which were used in the questionnaire we made extensive research by reading articles related to the topic of our project, in particular about “trust”.

After that we prepared the questions with the collaboration of the supervisor and sent out a hard copy version of the questionnaire for ten students to get a feedback mainly about the type of questions we used and their opinion about it. Based on the feedback we get from them we reduced the number of questions in the questionnaire which was originally 27 to 23, and we also made changes on the ways of some of the questions were presented. Then we translated the questionnaire to Norwegian language to give people an option to use one of the two languages which they are comfortable with. After that we created an account and sent out the questionnaire via [surveymonkey.com](https://www.surveymonkey.com), and distributed the link by using email and face book. The whole questionnaire can be seen in the appendix section.

After the data was gathered, the diagrams were created and analyzed with the help of [surveymonkey.com](https://www.surveymonkey.com).

Finally, we arranged a group of 4 programmers who were former students of this particular masters program to be part of the session. This method was chosen because it makes it possible to get an impulsive response from participants at that particular time and to put into perspective whole different ideas that can be proposed by discussing the matter as a group. At the beginning of the discussion I presented to the group the findings of the questionnaire, in order to give them an overview of the key findings. The discussion continued by raising some of the surprising findings of the questionnaire and the possible factors which made the participants to select them. The focus-group discussion was held in UIS.

3.2 Results of Survey

In this section we will explain the results we get from the questionnaire and the focus group. This questionnaire was sent out via surveymonkey.com and distributed to participants by face book and email; as a result a response from 108 participants was acquired.

The majority, 66 % of the participants was male and 34 % of the participants were female, as it's shown in the figure below. The average mean age of the participants was 27, with the youngest age 20 and the oldest 34.

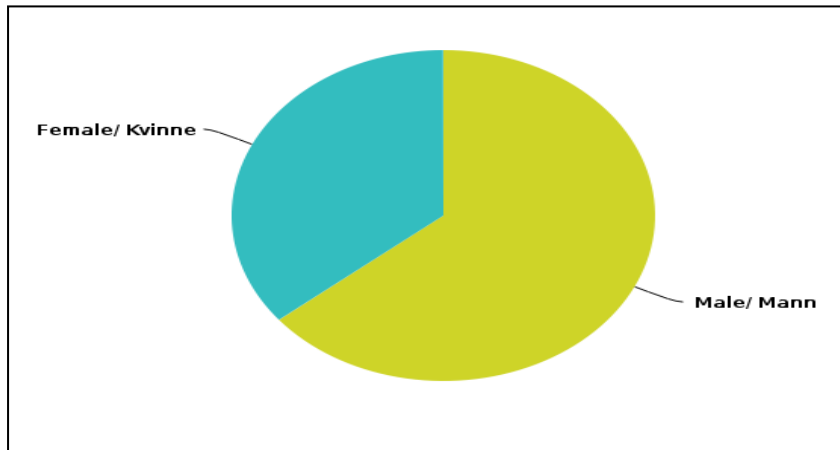


Figure 2- Percentage of female and male participants in the survey

Q. Are you part of a social network society? (Example - Face book, MySpace, tweeter)

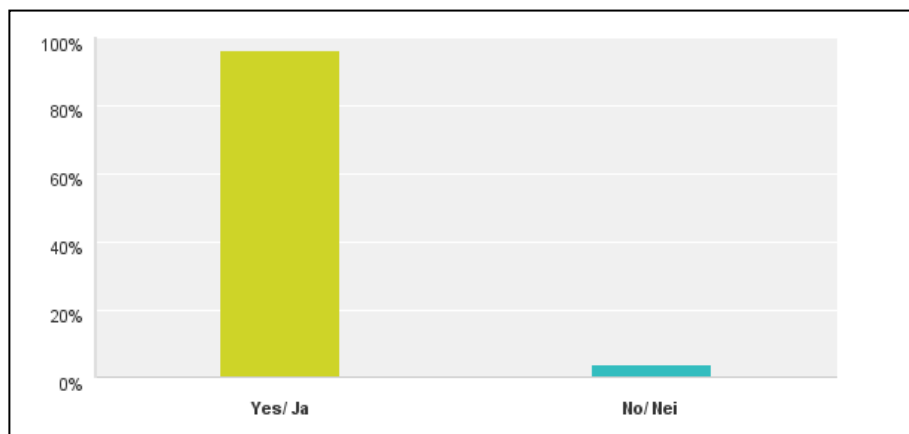


Figure 3 shows the percentage of social media network members

Key Findings

- 96 % of the participants said Yes
- the remaining 4 % said No

Q. Is the number of people who commented or like a link which is shared in social media important for you when it comes to trusting the information?

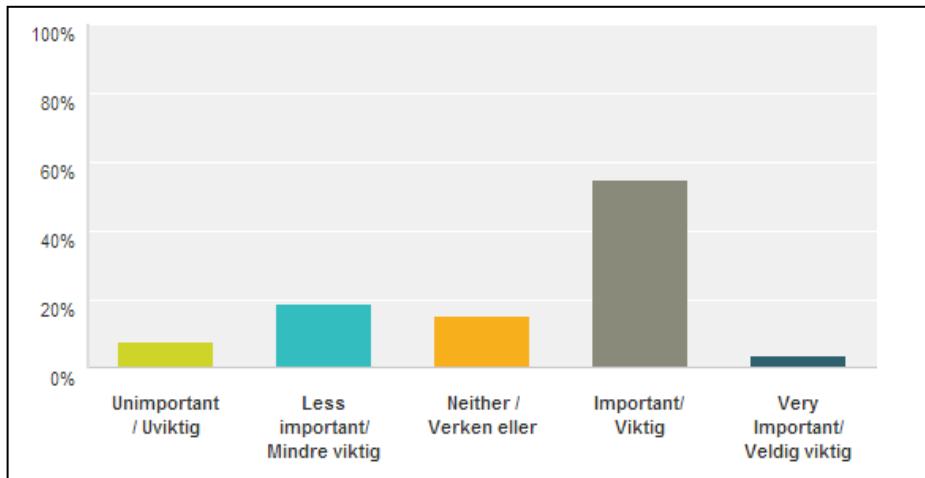


Figure 4 points out the importance of the number of people who commented or like a link

Key Findings

The participants were given the option to choose their answer from five categories, namely Very important, Important, Neither, Less Important and Un Important. 7,5 % said it is un important , 18,9 % said it is less important, 15,1 % said it is neither, 54,7 % said it is important and at last 3,8 % said it is very important.

Q. Knowing the person who shared the information (it could be personally) important for you?

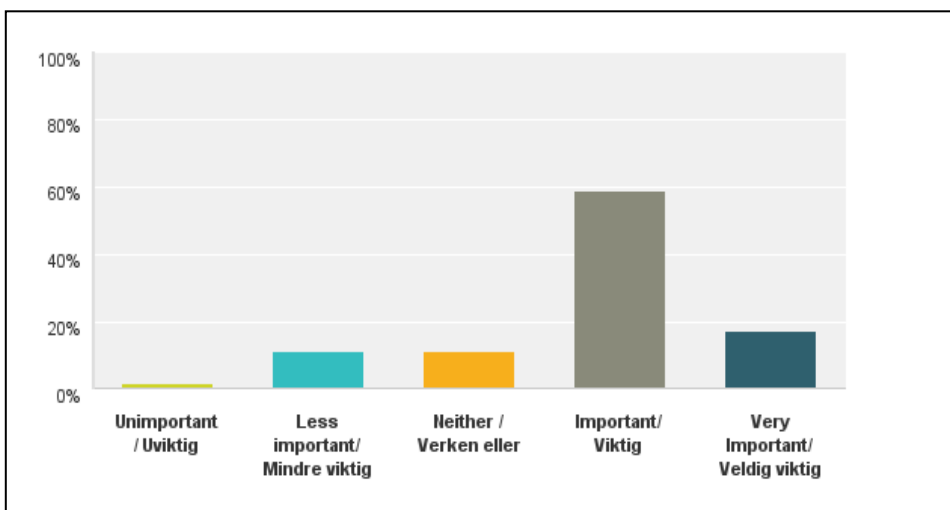


Figure 5 displays the factor of knowing the person who shared the information

Key Findings

17% of them said it is very important. While the majority, which is 58,5 % said it is important for them, 11,32 % of them said both it is Neither and less important respectively and only 1,9 % said it is un important.

Q. Do you think engaging actively in social media will make a person more trustworthy?

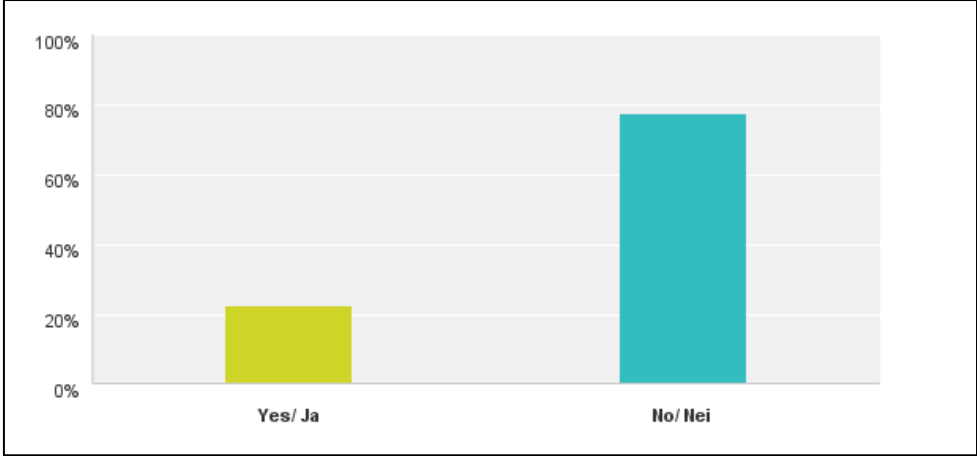


Figure 6 indicates how participants think about engaging actively in social media in relation to trust worthiness

Key Findings

- Only 22,6 % of the participants said yes
- 77,6 % who said it doesn't matter(No).

Q. Do you use more than one social media networks?

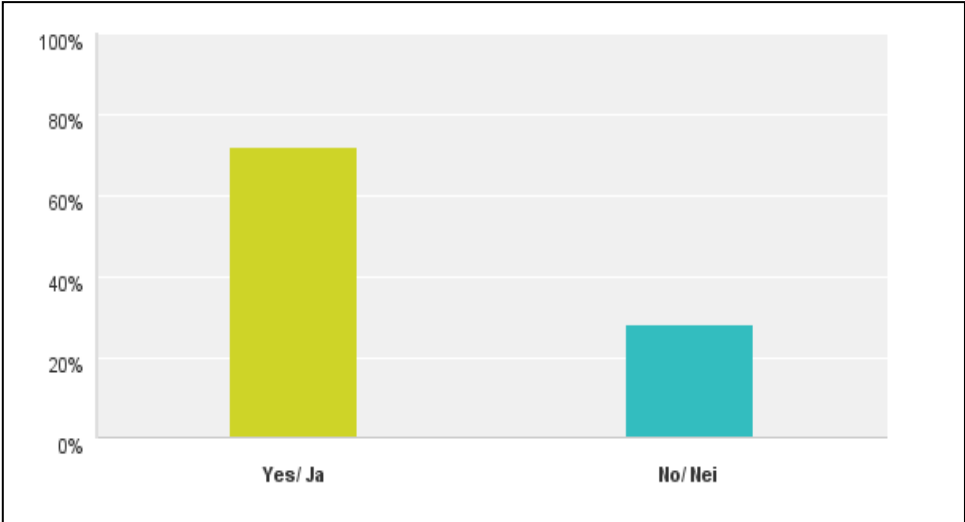


Figure 7 shows percentage of participants who use more than one of the available social network sites

Key findings

- 71,7 % said yes
- 28,3 % said No.

Q. In your opinion, how important it is for a person to increase his trustworthiness by being actively engaged in more than one social media networks ?

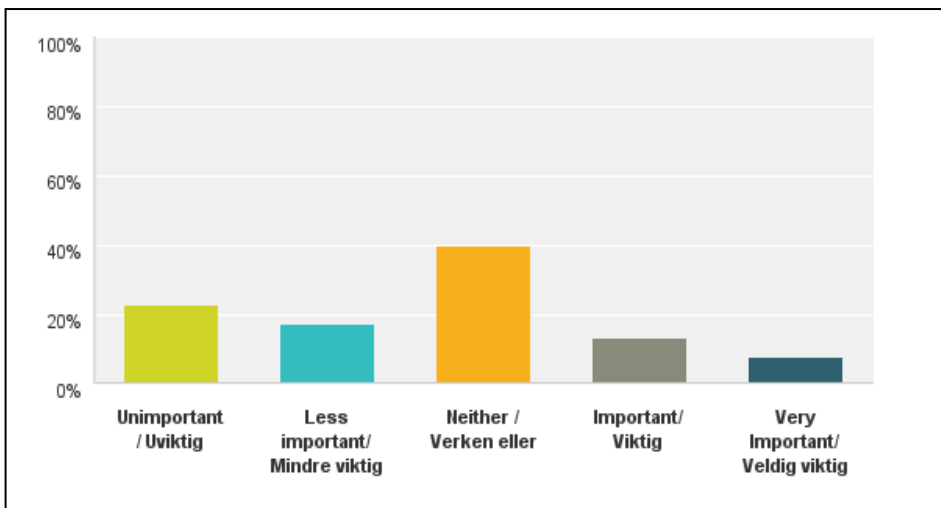


Figure 8 displays importance of engaging in more than one social network site when it comes to increasing trustworthiness

Key Findings

- Almost 40% of them said neither, 22 % said unimportant, 17 % said less important. The percentage of people who said it is important and very important is 13 and 7,5 respectively.

Q. Is the number followers or friends the person sharing the information have influences your assessment of the credibility of the content?

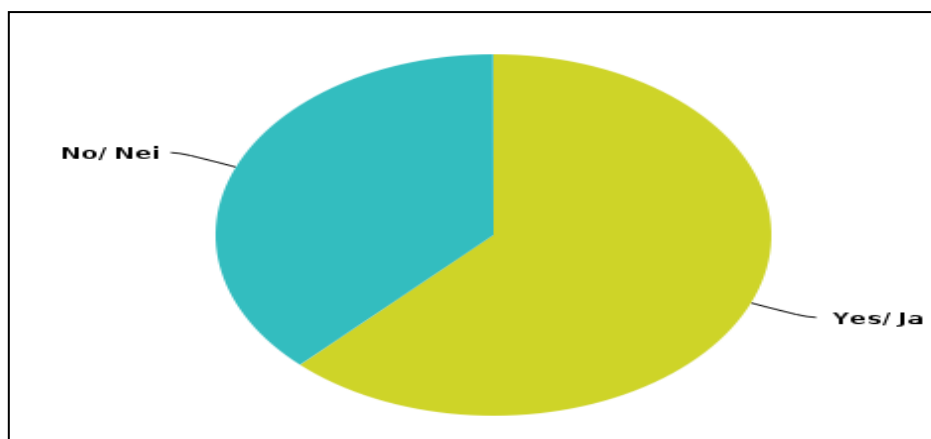


Figure 9 indicates people's opinion of social media users with both many or few friends and the credibility of the information they post

Key findings

- Majority of the participants (62,25 %) said yes
- 37,75 % said no.

Q. Does the trustworthiness of a person depends on the quality of the previous posts, comments and links he/she shares?

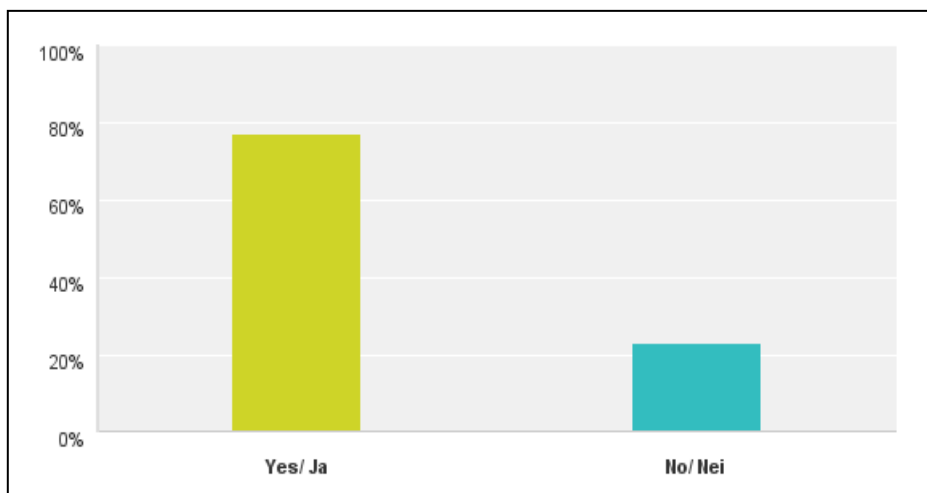


Figure 10 shows the views of participants on a previous posts quality importance for trusting the future posts

Key Findings

Here 77 % of the participants answered yes and the other 23 % said no. The result shows that if the person have a record of sharing un confirmed information which happen to be incorrect often in the past, it will make it quite hard for the information he will share in the future to be trusted by his friends and the same goes with a person who have a previous record of sharing accurate information.

Q. On average, how many people should share a content before you start trusting the information?

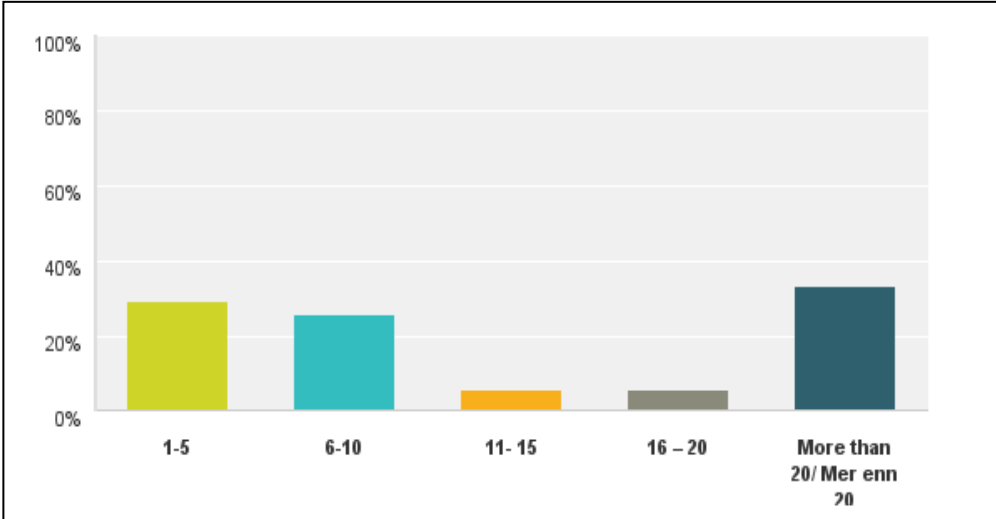


Figure 11 shows how many times a post should be shared for participants to start trusting it

Key findings

Here 29,4 % think that 1-5 is enough, 25,6 % think 6 - 10, 6 % think 11 - 15, another 6 % think 16 - 20 and 33 % think more than 20 is necessary to start trusting the information.

Q. Do you think the information which is shared in social media is higher quality (trust worthy) than the traditional media outlets such as television, radio and newspapers?

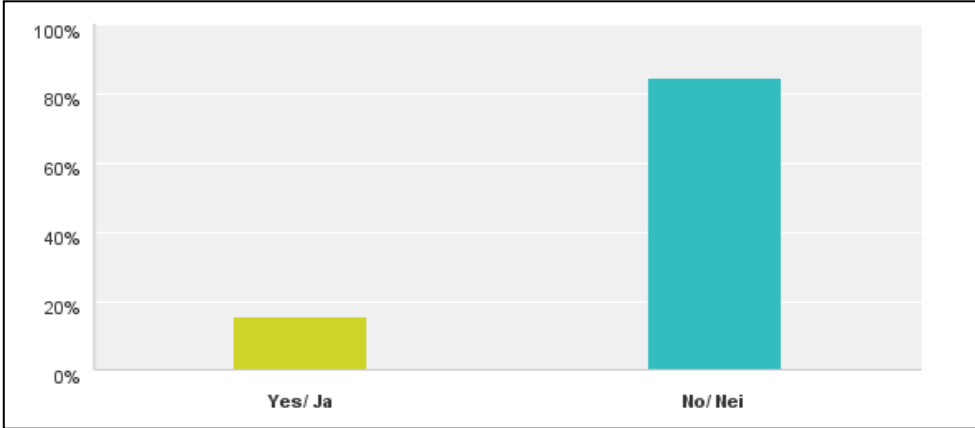


Figure 12 display participants trust in traditional Vs social media

Key Findings

- Only 15,4 % said yes
- and the overwhelming majority which is 84,6 % said no.

Q. Which social media platform is your favorite?

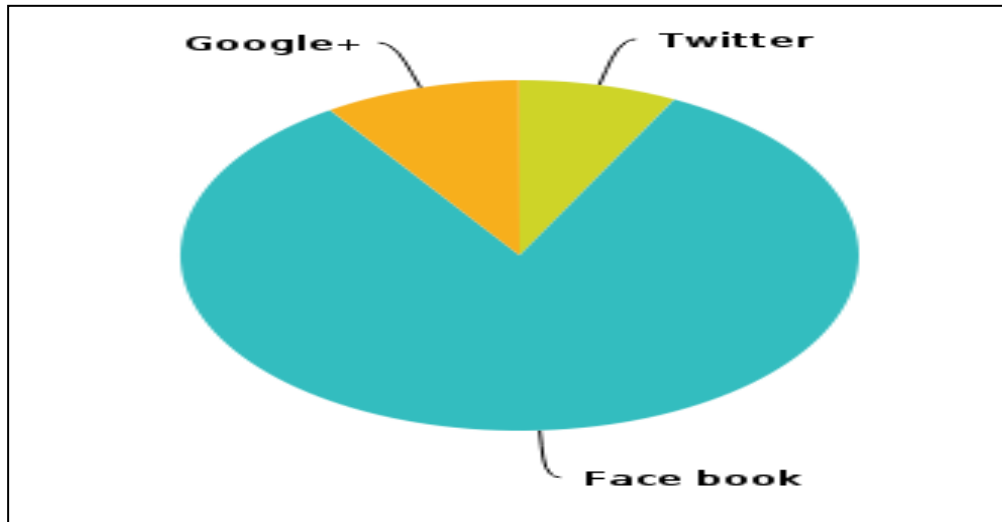


Figure 13 shows the percentage of participants favorite social media sites

Key Findings

The clear favorite was face book with 83 %, followed by Google + with 9,4 % and Twitter with 7,6 %.

Q. Have you ever blocked or “unfriended “ people from your friends list because of the untrustworthiness of the information they share?

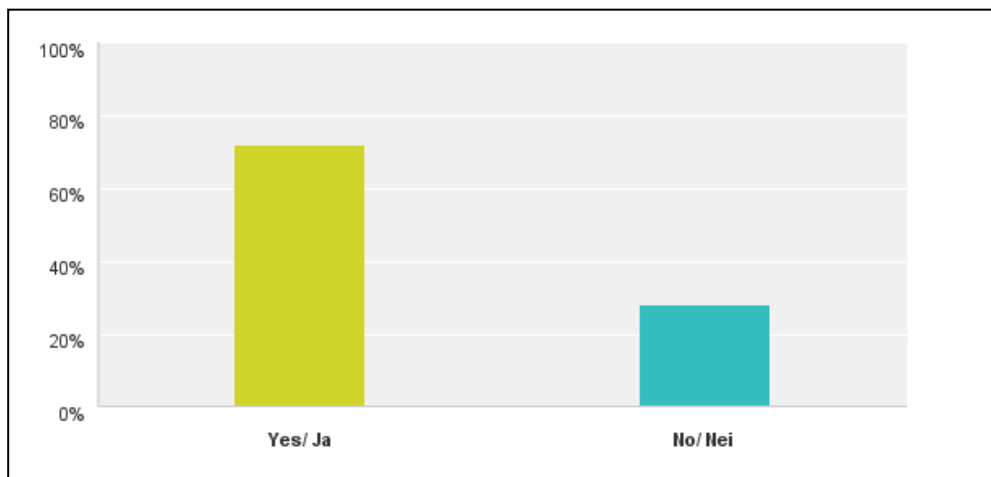


Figure 14 indicates how many of the participants blocked or un-follow(in case of tweeter) people due to the fact that information they share is often inaccurate

Key Findings

- 28,3 % said no
- the other 70 % said Yes.

Q. Which of the following is your most important news source?

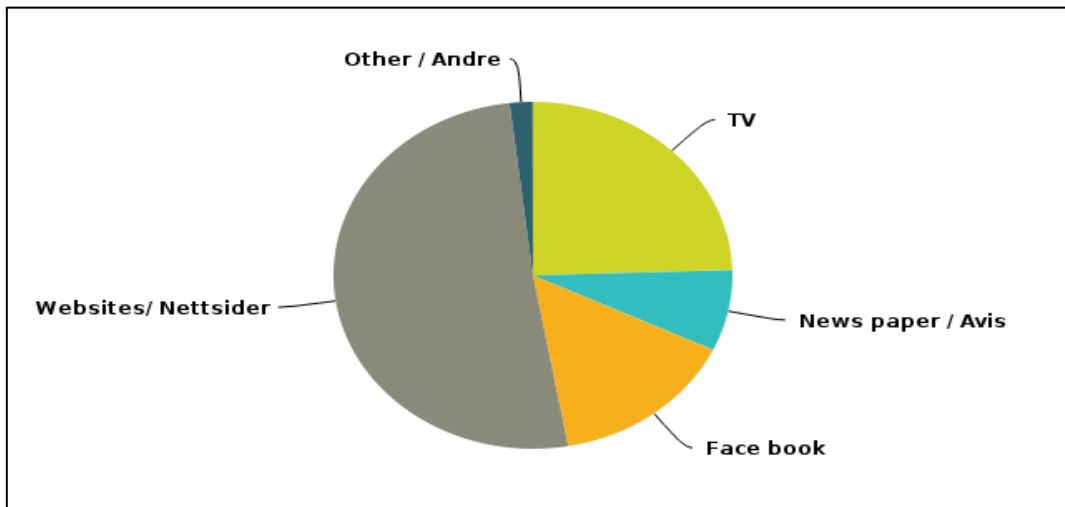


Figure 15 shows the news sources which are popular among the participants

Key findings

- Obviously 51 % said websites
- followed by 24.4 % TV
- 15.1 % Face book , 7,6 news paper and 1,9 % said others
- Surprisingly the result for tweeter was 0 %.

Q. How much trust do you have in social media as a source of news? In a scale of 0 to 5 (5 if you fully trust them and 0 if you don't trust them at all).

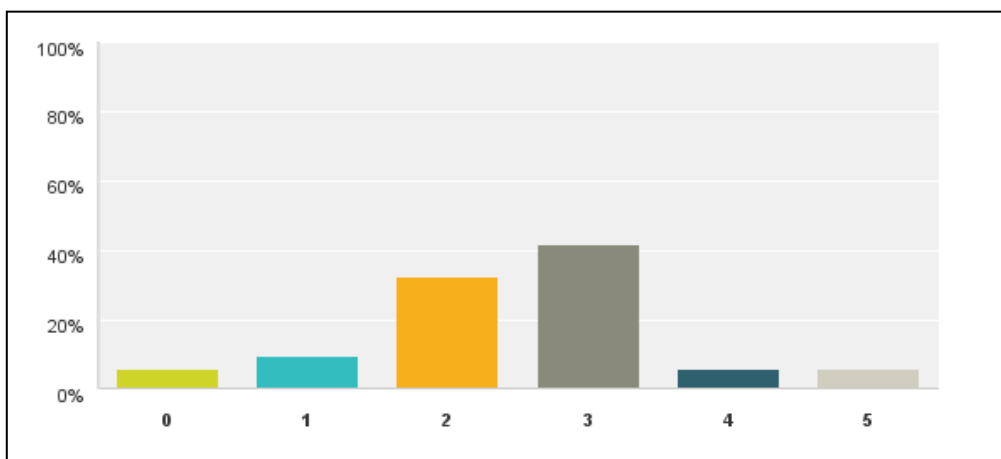


Figure 16 shows the rating of social media as news sources

Key Findings

And results were 5,66 % said 0 , next 9,43 % said 1, followed by 32 % said 2 , 41,5 % said 3 and 5,66 % each for 4 and 5 .

Q. Which of the following do you need to trust to a social media content? (you can select multiple) Please also order these criteria from the most important to the least.

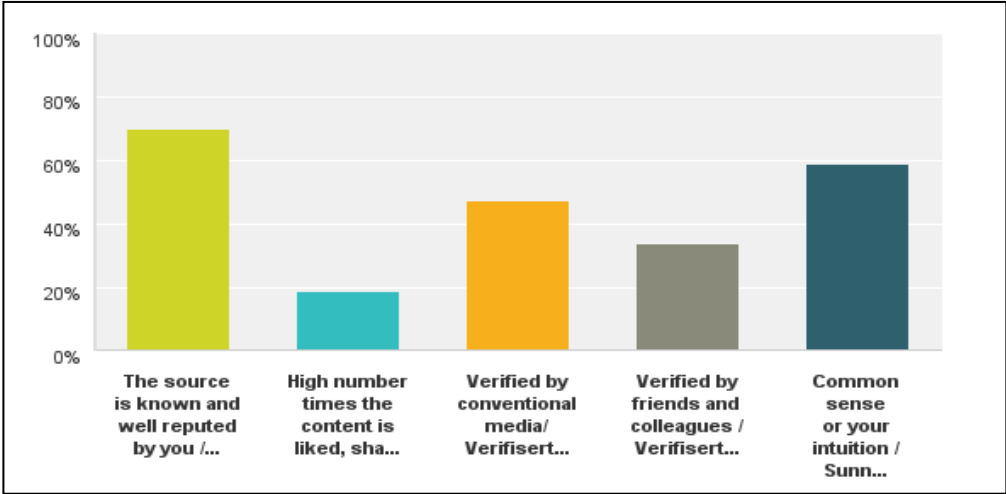


Figure 17 displays participants opinion about the reasons which make them to trust a social media content

Key Findings

- The source is known and well reputed by you , 69 %
- High number times the content is liked, shared and forwarded , 17 %
- Verified by conventional media, 47 %
- Verified by friends and colleagues , 33 %
- Common sense or your intuition, 58 %

Q. Which of the following make you NOT trust to social media content? (You can select multiple) Please also order these criteria from the most important to the least.

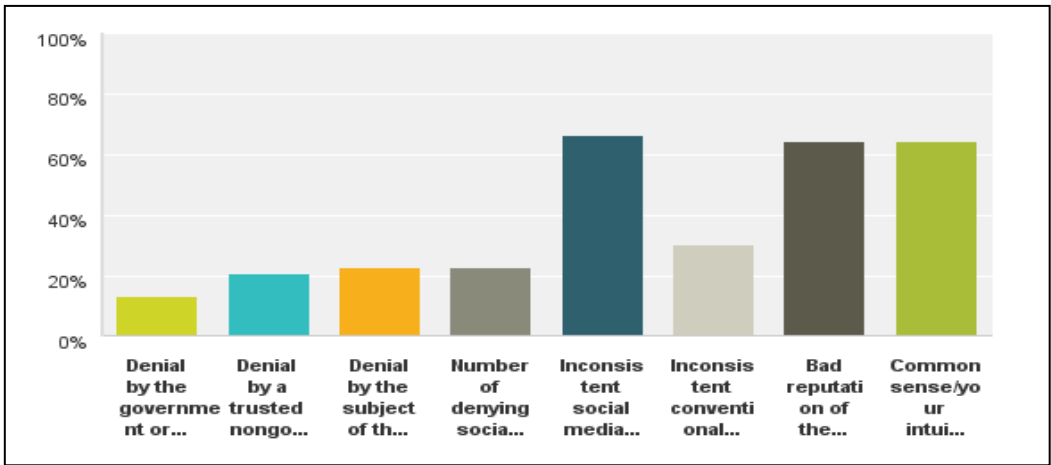


Figure 18 indicates participants opinion about the reasons which make them NOT to trust a social media content

Key Findings

- 1- Denial by the government or a governmental organization, 13,21 %
- 2- Denial by a trusted nongovernmental organization, 20,75 %
- 3- Denial by the subject of the content, 22,64 %
- 4- Number of denying social media content, 22,64 %
- 5- Inconsistent social media content, 66,04 %
- 6- Inconsistent conventional media content, 30,19 %
- 7- Bad reputation of the source, 64,15 %
- 8- Common sense/your intuition, 64,15 %

At last we will analyze the open-ended questions of the questionnaire. In this questionnaire we have included five open ended- questions excluding the question which ask the participants age. We will go through them sequentially like the way they are ordered in the questionnaire. The first open-ended question inquires for the number of years the participants used social network sites. The responses vary from a minimum of one year to the maximum of eleven years, but the majority of users response was 5 years. The main reason behind that was rapidly increasing popularity of face book and twitter at that moment.

The second one was about the participant's field of study, which was very diverse. To mention some of the areas of studies – Medicine, Teacher, Bio chemistry, protein chemistry, Economics, Computer science, Pharmacy, Social works etc...

After that the third open-ended question presented to participants were asked if they share or forward any information which they don't fully trust, almost all the participants replied no with exception of only two who replied some times.

Next participants were asked if they have any other criteria which they need to trust a social media contest which is different from the one proposed in the previous questions. Here some of them propose some newer ideas such as the quality that the information is presented tends to effect my tendency to take it take it seriously, scientifically proven if possible, should be reported by freelancers with out any political party affiliation, trust worthiness of the people who shared it with me, if it doesn't have inconsistencies, or vague references/reasoning or unsupported claims and so on.

And finally they were inquired if they have any other criteria that make them NOT trust to a social media content and most of the replies were pretty much the same with the earlier question replies.

3.3 Focus Group Result

Professional people's with computer science or information technology educational background opinions towards the trust issues is very important in addition to the survey which was conducted randomly on people with different educational backgrounds.

For this reason, I arranged a group of 4 programmers who were former students of this particular masters program to be part of the session. This method was chosen because it makes it possible to get an impulsive response from participants at that particular time and to put into perspective whole different ideas that can be proposed by discussing the matter as a group.

As it is mentioned above, 4 people were participated in the discussion

Participants of Focus- group		
<i>Age</i>	<i>Gender</i>	<i>Occupation</i>
32	Male	IT Consultant
29	Male	Fellow Research (UIS)
30	Male	Software Developer
27	Male	Software Developer

Table.1 Participants of a focus group.

At the beginning of the session I presented to the group the findings of the questionnaire, in order to give them an overview of the key findings. As it is shown in the result of the questionnaire, majority of the participants said that being actively engaging actively in a social media doesn't increase his trustworthiness. So, what else should a person have to do to get the trust of other people? beside engaging actively, was the first question we began the discussion with.

Then, one of the focus group discussion members stated his opinion by saying

“May be most of the participants of the questionnaire come to this conclusion because they didn't consider it in marketing perspective, instead only in personal perspective. Imagine if you are on a company page which doesn't address client complaints in time and which doesn't update it's status regularly even though it is getting many negative reviews, and in contrast there is a company which respond to the critics regularly and said thank you for those who are praising the products of the company. The one that is responding regularly will definitely gain more trust from the people who are reading the debates and conversations.”

Then another participant joined the discussion by saying

“if a person or let's said firm is actively participating in social media , it shows that person is accountable. If we take marketing by using social media as an example, when people perceive that they can communicate with a person in a meaningful discussion whenever they want knowing that they will get a reply instantly, the level of trust they have towards the person or brand will increase. “

Then the debate continues on another finding of the questionnaire which was information shared in traditional media outlets have seen as a higher quality than social media outlets. One of the participants said that

“ I think the main reason is that, social media lacks any accountability when it comes to fact checking and accuracy of a content unlike TV, newspaper and radio. This affects significantly its trustworthiness. “

“Social media outlets clearly depend on the news from traditional media outlets (mainstream medias) to a large extent, since they don't have their own journalists”

So, why do you think then when asked about their important news source the majority of the participants said websites, even though TV (Traditional media) becomes their second most important news source by a quite big margin?

“This days more people are turning their way into social media to keep in touch with everyday news. Nowadays nearly all of the traditional media outlets have their presence in social network sites. Traditional media outlets doesn’t see social media sites as a competitors, instead they see them as a means which helps them to distribute their content”

“For me, when I want to have interaction, collaboration or the other amazing features that social media offers, I usually go to social network sites. However if my aim is to get news, particularly news about areas which I don’t have a thorough understanding or a clue, I would prefer traditional media outlets with journalists who have a good know-how of the areas.”

At last, participants in the focus group made extensive reviews of the way the social network sites are designed and what can be done to improve or simplify their design in a way that could help the ordinary user to differentiate easily whether a content could be trusted or not. Even though those discussions were beyond the scope of this particular project and not explained here, they will be used when we start working on the master thesis. Then participants wished me a good luck in my project and told me their willingness to participate on future focus group discussion sessions if they are needed, by that we end the session.

3.4 Selected attributes

No.	Parameter Name	Description	Data Type
1	Age	The age of the participant from the survey.	Numeric
2	Years of use	The number of years the particular participant used a social media.	Numeric
3	Gender	Gender of the participant	Nominal
4	Number of followers	Whether number of followers the person who is sharing the content matters or not.	Nominal
5	Forwarding un trusted content	If they ever forward or repost an untrusted(unconfirmed information) on social media	Nominal
6	Number of likes	If the number of people who commented or like a link which is shared in social media important when it comes to trusting the information	Nominal
7	Important News Source	Which of the following is your most important news source	Nominal
8	Number of people Sharing	how many people should share a content before you start trusting the information	Numeric
9	Social Vs Traditional Media	Do you think the information which is shared in social media is higher quality (trust worthy) than the traditional media outlets such as television, radio and newspapers	Nominal
10	Using > 1 social media	The importance for a person to increase his/her trustworthiness by being actively	Nominal

		engaged in more than one social media networks	
11	Blocking a person	Have you ever blocked or “unfriended “ people from your friends list because of the untrustworthiness of the information they share	Nominal
12	Favorite social network	Which social media platform is your favorite	Nominal
13	Trust in previous posts	If the trustworthiness of a person depends on the quality of the previous posts, comments and links he/she shares	Nominal
14	Field of Study	The type of education the participant have	Nominal
15	Trust in SN	The trust you have in social media as a source of news(scale of 0 to 5)	Numeric

Table – description of the selected attributes

All the attributes were assigned a numeric values as it is shown below

1. Age: the age of the participant from the survey.
2. Gender: 0 = MALE, 1 = FEMALE
3. Years of use: the number of years the particular participant used a social media in numerical form.
4. Number of followers: 0 = YES, 1 = NO
5. Forwarding un trusted content: 0 = YES, 1 = NO
6. Important News Source: 0 = Websites, 1 = Face book, 2 = News paper, 3 = TV and 4 = Others
7. Number of people Sharing: More than 20 = 5, 15-20 = 4, 10-15 = 3, 5-10 = 2 and 1-5 = 1.
8. Social Vs Traditional Media: 0 = YES, 1 = NO.
9. Using > 1 social media: 0 = YES, 1 = NO
10. Blocking a person: 0 = YES, 1 = NO
11. Favorite social network: 0 = Face book, 1 = Tweeter 2 = for Google+
12. Trust in previous posts: 0 = YES, 1 = NO
13. Field of Study: 0 = Natural science fields and 1 = Social science fields.
14. Trust in SN: 0 = 0-2, 1 = 3-5

Relation: FPSnr3												
2: Age	3: Years of use	4: nr of people sharing	5: Favourite SN in nr.	6: Imp.News Sr. Nr	7: Forwarding untrusted sr. In nr	8: S vs T media in nr	9: Blocking a pr. In nr	10: Trust in previous posts In nr	11: Use > 1 SN in nr	12: nr of followers in nr.	13: Field of s	Num
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nume
24.0	4.0	5.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0		
31.0	7.0	2.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0		
28.0	3.0	2.0	1.0	2.0	0.0	1.0	0.0	1.0	0.0	1.0		
26.0	5.0	2.0	0.0	2.0	0.0	0.0	1.0	1.0	1.0	1.0		
28.0	11.0	5.0	2.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0		
21.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		
30.0	4.0	5.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0		
25.0	4.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0		
27.0	9.0	5.0	0.0	3.0	0.0	0.0	1.0	1.0	1.0	1.0		
24.0	5.0	5.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0		
29.0	7.0	5.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0		
26.0	4.0	4.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0		
31.0	8.0	2.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0		
20.0	5.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0		
28.0	5.0	1.0	0.0	2.0	0.0	0.0	1.0	1.0	1.0	0.0		
26.0	3.0	5.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0		
29.0	10.0	1.0	0.0	2.0	0.0	0.0	1.0	1.0	1.0	0.0		
30.0	4.0	5.0	2.0	2.0	0.0	0.0	0.0	1.0	1.0	1.0		
29.0	5.0	3.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0		
27.0	7.0	2.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0		
29.0	5.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		
34.0	9.0	2.0	0.0	2.0	0.0	0.0	1.0	1.0	0.0	1.0		
25.0	4.0	3.0	0.0	2.0	0.0	0.0	1.0	1.0	1.0	1.0		
27.0	5.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0		
30.0	6.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0		
29.0	5.0	1.0	0.0	3.0	0.0	1.0	1.0	0.0	1.0	1.0		
25.0	5.0	3.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0		
24.0	7.0	2.0	0.0	2.0	0.0	0.0	1.0	0.0	1.0	1.0		
22.0	9.0	5.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0		
29.0	4.0	2.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0		
29.0	4.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
28.0	10.0	2.0	0.0	3.0	1.0	1.0	1.0	1.0	1.0	1.0		
30.0	9.0	5.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0		
23.0	11.0	5.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0		
25.0	5.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0		
30.0	6.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0		
23.0	5.0	5.0	0.0	2.0	1.0	0.0	1.0	1.0	1.0	0.0		

Figure 20 - Snapshot of the preprocessed data

CHAPTER FOUR

4.1 Experimentations

Eleven experiments were carried out in total for this research. This chapter explains all the steps and procedures which happened during the experimentations. As it's described in the previous chapters, the objective of this thesis is, to discover patterns to predict people whether they trust or distrust a particular content with in the trust database. To accomplish our goal, the model-building phase in the DM process of this investigation was done using a three-step process. Clustering, classification and Regression were used in a subsequent order.

The K-means algorithm (using two different types of distance functions and four different seed values) was chosen to deal with clustering task of data into the two target classes of trust and distrust. Then, classification was performed to predict trust for each participant. The training data set was used when dealing with both clustering and classification processes, and both tasks were performed using Weka 3.6.11 DM tool.

Subsequent to conducting the experiments, the models were evaluated using different performance measures like time span, accuracy, TP Rate, FP Rate, F-Measure and ROC Area. This research also conducts experiments on linear regression, logistic regression, Poisson regression and negative binomial regression within the survey data. After comparing the above mentioned models on the basis of AIC, log-likelihood and the two deviances(null and residual) the best alternative model will be selected.

4.2 Cluster Modeling

Four experiments were carried out for the purpose of building a cluster model for this research, by changing the different parameters of the K-Means Algorithm. All four experiments will be explained in detail and their respective output will be analyzed. Finally, we will compare the output of the four experiments based on their values of number of iteration, within cluster sum of squared errors and the time it took to build the model. After that the best cluster model will be selected and to construct our final model.

In our experiments we split the full training set into two and then we allocate 75 % of the data set for training and the remaining 25% for the purpose of testing data set.

While doing the experiments in WEKA's K-Means clustering, there are certain parameters we have to change for each experiment. Some of those parameters are explained below

Explanation	Name of the Parameter	Usage
-------------	-----------------------	-------

A function which is used to calculate the distance	Distance function	To select the type of distance function to be used
The number of clusters	K	To assign the K value
The number of data tuples the cluster should start with	Seed Value	To assign a random seed value

Table– The parameters used in the experimentation with their explanation

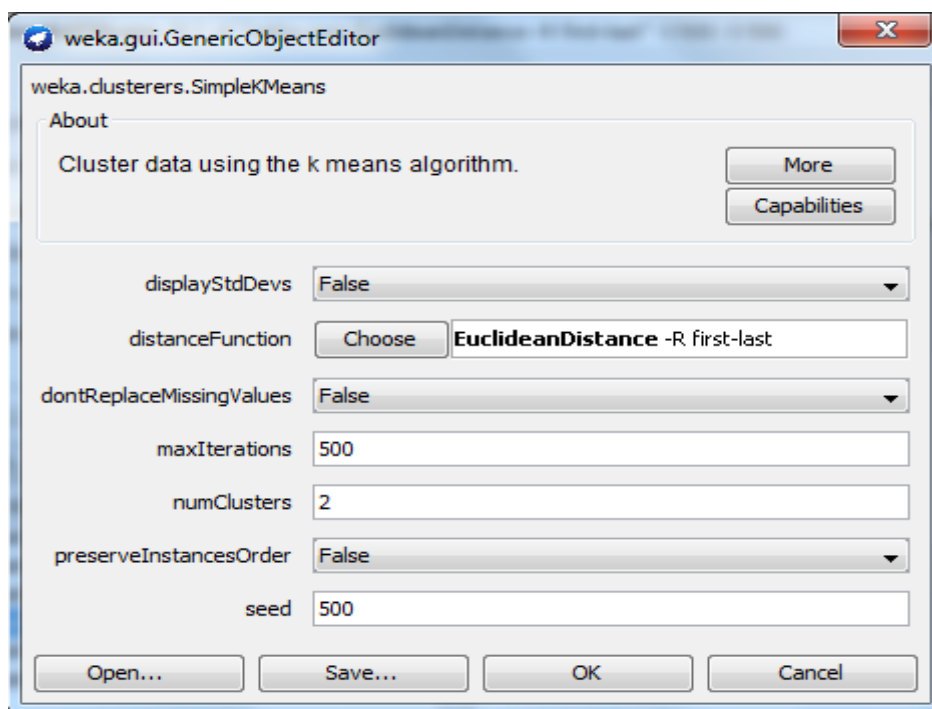


Figure 21 - Cluster attributes

4.2.1 Experiment 1

This experiment was performed for $K=2$, with default values of seed and distance function. Every one of the final chosen 14 attributes and 106 records were used in this experiment. For the purpose of clustering the records according to their values this model was trained by using the default values of the K-Means algorithm. The table below shows the outcome of the experiment and cluster distribution of the data set.

Cluster Result				
Distance Function	Seed Value	K	Cluster Distribution	
			C0	C1
Euclidean Distance	10	2	45(42%)	61(58%)

Table – The values of the parameters used for the first experiment

According to the above table, we can clearly observe that the first experiment was performed with default values of the algorithm (Euclidean distance, K = 2 and Seed Value= 10).

Attribute	Full Data (106)	Cluster#	
		0 (45)	1 (61)
Gender in nr.	0.6415	0.4667	0.7705
Age	27.2075	27.0444	27.3279
Years of use	5.6792	5.6444	5.7049
nr of people sharing	2.8302	2.6444	2.9672
Favourite SN in nr.	0.2642	0.3333	0.2131
Imp.News Sr. Nr	0.9623	1.0222	0.918
Forwarding untrusted sr. In nr	0.066	0	0.1148
S vs T media in nr	0.1792	0.2667	0.1148
Blocking a pr. In nr	0.6981	0.4667	0.8689
Trust in previous posts In nr	0.7642	0.4667	0.9836
Use > 1 SN in nr	0.717	0.5111	0.8689
nr of followers in nr.	0.6226	0.8	0.4918
Field of study in nr.	0.3774	0.3556	0.3934
Trust in SN binary	0.5283	0.6889	0.4098

Figure 22 - Clustering output of the first experiment

The output is showing us the togetherness of the clusters, "1" means all of them in that cluster share the exact same value of one, and a "0" means all of them in that cluster has a value of zero for that particular attribute. The other numbers are mostly the average value within in the clusters. Individual clusters exhibits a type of behavior in our participants, based on which we can start to draw some conclusions.

Because this experiment has created a bigger number of distrust claims (61), in comparison to trust claims of 45 performing other experiments becomes quite necessary. Moreover, the output of the experiment exhibits us that within cluster sum of squared error is a little bit high, which leads to the fact that instances within the same cluster have a tendency to not have similarity. In order to improve this result the next experiment was done with a seed value of 100.

Another way of inspecting the data in these clusters is to observe it visually. As it is shown in the diagram below, by changing the X and Y axis's to each and every one of our attributes it is possible to observe clearly the way the clusters are grouped and organized.

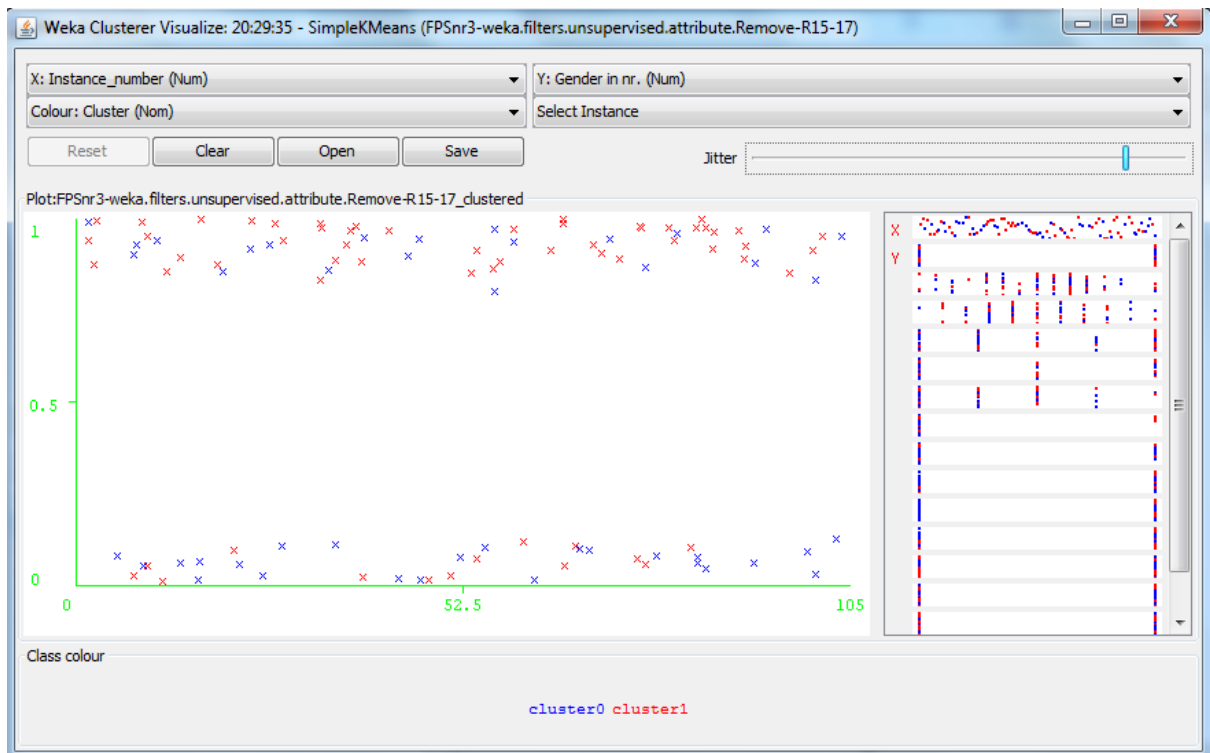


Figure 23- Cluster visual inspection of first experiment

4.2.2 Experiment 2

The second experiment was carried out with a default K value, a default distance function (Euclidean Distance) and seed value of 50.

Cluster Result				
Distance Function	Seed Value	K	Cluster Distribution	
			C0	C1
Euclidean Distance	50	2	74(70%)	32(30%)

Table – The values of the parameters used for the second experiment

The Figure below shows the results of the second experiment.

Attribute	Full Data (106)	Cluster#	
		0 (74)	1 (32)
Gender in nr.	0.6415	0.6757	0.5625
Age	27.2075	27.3378	26.9063
Years of use	5.6792	5.7297	5.5625
nr of people sharing	2.8302	2.973	2.5
Favourite SN in nr.	0.2642	0.2162	0.375
Imp.News Sr. Nr	0.9623	1	0.875
Forwarding untrusted sr. In nr	0.066	0.0541	0.0938
S vs T media in nr	0.1792	0.1892	0.1563
Blocking a pr. In nr	0.6981	1	0
Trust in previous posts In nr	0.7642	0.8108	0.6563
Use > 1 SN in nr	0.717	0.7973	0.5313
nr of followers in nr.	0.6226	0.6081	0.6563
Field of study in nr.	0.3774	0.3378	0.4688
Trust in SN binary	0.5283	0.5405	0.5

Figure 24 - Clustering output of the second experiment

As in the first experiment, the result is showing us the togetherness of the clusters, "1" means all of them in that cluster share the exact same value of one, and a "0" means all of them in that cluster has a value of zero for that particular attribute. The other numbers are mostly the average value within in the clusters. Individual clusters exhibits a type of behavior in our participants, based on which we can start to draw some conclusions. In addition, we can observe each cluster visually in the same manner as it's explained in the first experiment.

This experiment gives a much improved result in comparison with the first experimentation, the value of within clustered sum of squared error is minimized to 207.58 and also the number of iteration that the K-Means algorithm used to converge was also lowered from 7 to 5. Moreover, the number of trust claims 70% (74) was also higher than the distrust claims 30% (32) in this experiment.

The result of this experiment looks quite satisfactory, however performing other experiments by changing the type of distance function and seed values seems quite important in case we find much better clustering model.

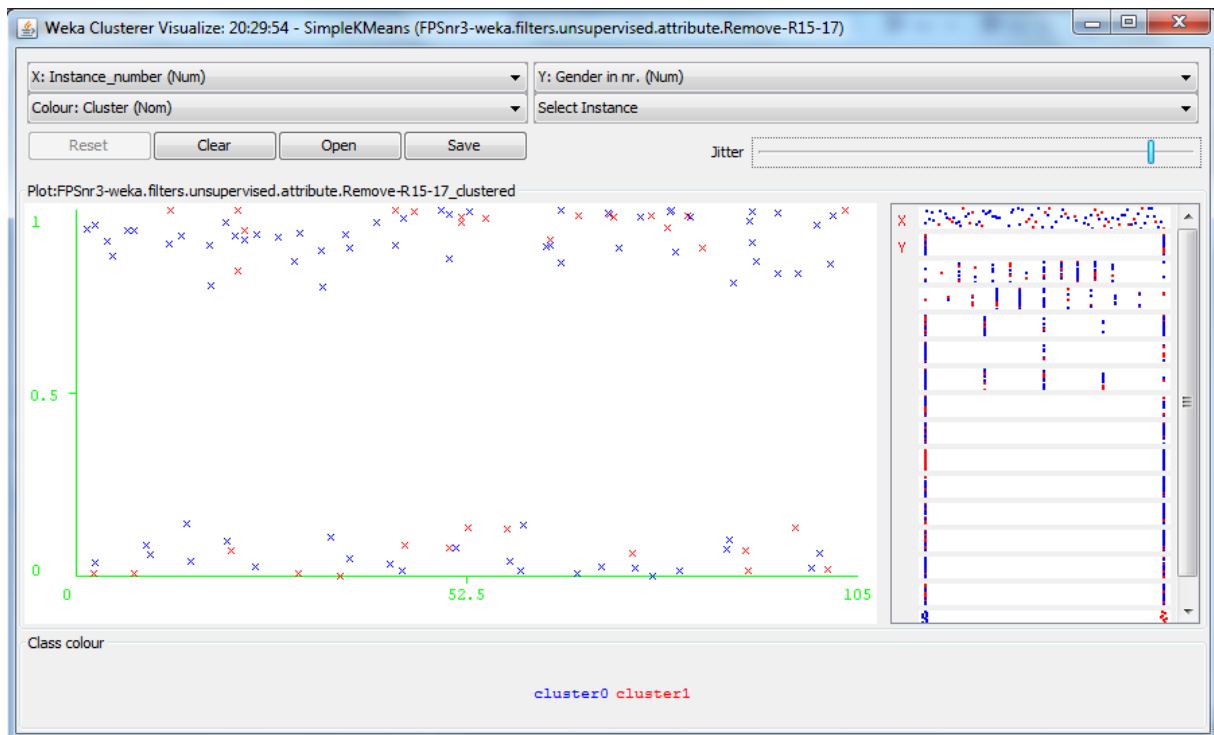


Figure 25 - Cluster visual inspection of second experiment

4.2.3 Experiment 3

The third experiment was performed with a seed value of 100, while K and Distance Function keep the default value. The table below exhibits us the parameters used in this experiment and the segmentation of individual clusters.

Cluster Result				
Distance Function	Seed Value	K	Cluster Distribution	
			C0	C1
Euclidean Distance	100	2	51(48%)	55(52%)

Table – The values of the parameters used for the third experiment

This experiment didn't give us a better result in comparison with the preceding two experiments, the value of within clustered sum of squared error increased to 208.11 and also the number of iteration that the K-Means algorithm used to converge was also maximized by 4 to become 9.

In addition, the number of trust claims 48% (51) was also lower than the distrust claims 52% (55) in this experiment, which definitely is not a good sign.

Attribute	Full Data (106)	Cluster#	
		0 (51)	1 (55)
Gender in nr.	0.6415	0.7059	0.5818
Age	27.2075	27.2941	27.1273
Years of use	5.6792	5.4314	5.9091
nr of people sharing	2.8302	2.5098	3.1273
Favourite SN in nr.	0.2642	0.2353	0.2909
Imp.News Sr. Nr	0.9623	1	0.9273
Forwarding untrusted sr. In nr	0.066	0.0392	0.0909
S vs T media in nr	0.1792	0.2745	0.0909
Blocking a pr. In nr	0.6981	0.6078	0.7818
Trust in previous posts In nr	0.7642	0.6275	0.8909
Use > 1 SN in nr	0.717	0.5098	0.9091
nr of followers in nr.	0.6226	0.6471	0.6
Field of study in nr.	0.3774	0.098	0.6364
Trust in SN binary	0.5283	0.8235	0.2545

Figure - Clustering output of the third experiment

In the diagram below, we can observe each cluster visually in the same manner as it's explained in the preceding experiments.

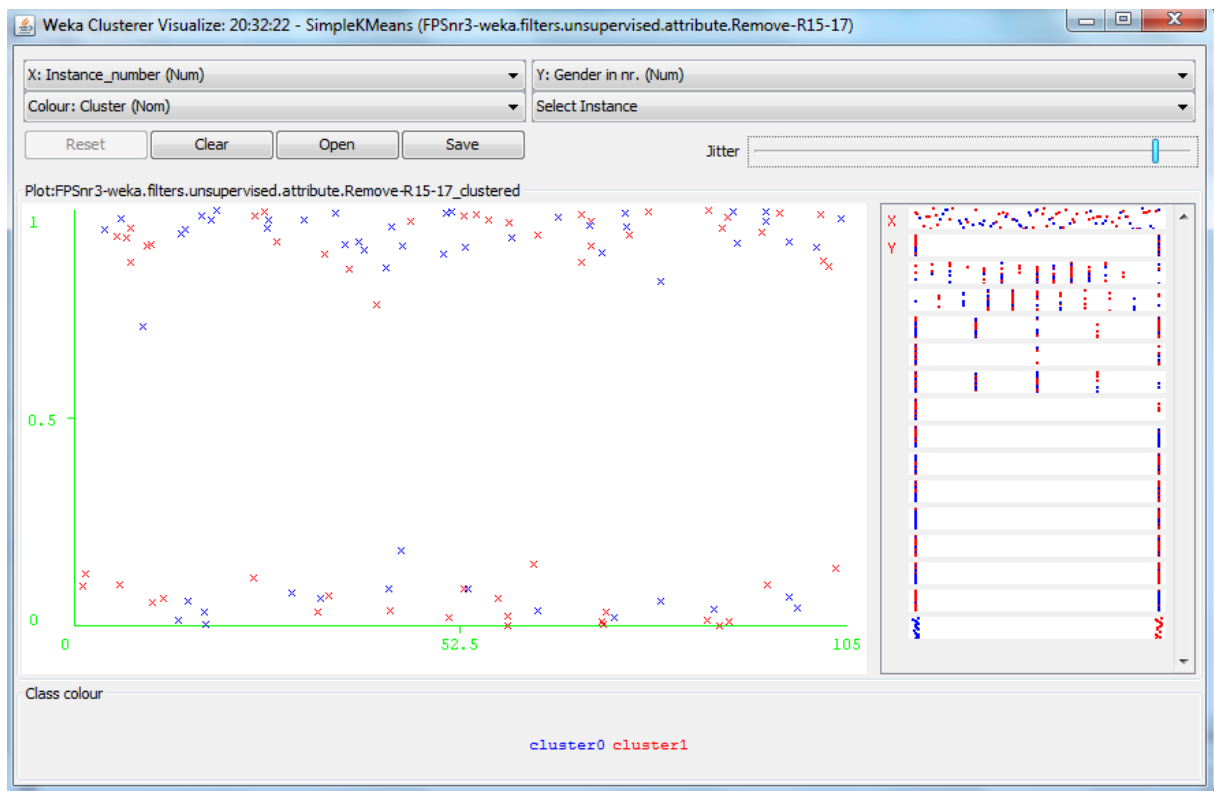


Figure 26 - Cluster visual inspection of third experiment

4.2.4 Experiment 4

Our final experiment was performed for $K = 2$, a seed value of 1000 and a new distance function by the name Manhattan Distance. Like the previous three runs every one of the (14) final chosen attributes and 106 records were used to carry out the experiment.

The table below shows the result of our final cluster experiment.

Cluster Result				
Distance Function	Seed Value	K	Cluster Distribution	
			C0	C1
Manhattan Distance	1000	2	53(50%)	53(50%)

Table – The values of the parameters used for the fourth experiment

Attribute	Full Data (106)	Cluster#	
		0 (53)	1 (53)
Gender in nr.	1	1	1
Age	28	28	28
Years of use	5	5	5
nr of people sharing	2	2	2
Favourite SN in nr.	0	0	0
Imp.News Sr. Nr	0.5	2	0
Forwarding untrusted sr. In nr	0	0	0
S vs T media in nr	0	0	0
Blocking a pr. In nr	1	1	1
Trust in previous posts In nr	1	1	1
Use > 1 SN in nr	1	1	1
nr of followers in nr.	1	1	0
Field of study in nr.	0	0	0
Trust in SN binary	1	0	1

Figure 27 - Clustering output of the fourth experiment

Even if this final experiment was performed with a new distance function (Manhattan distance function) and also a new seed value, the resulting cluster was not better than those of the previous three experimentations. Even though the number of iteration it took to converge was the smallest which is 3, the amount of within cluster sum of squared error was by far the highest in comparison with the preceding experimentations (353). This means, the experiment didn't manage to create is failed to create distinct clusters of trust.

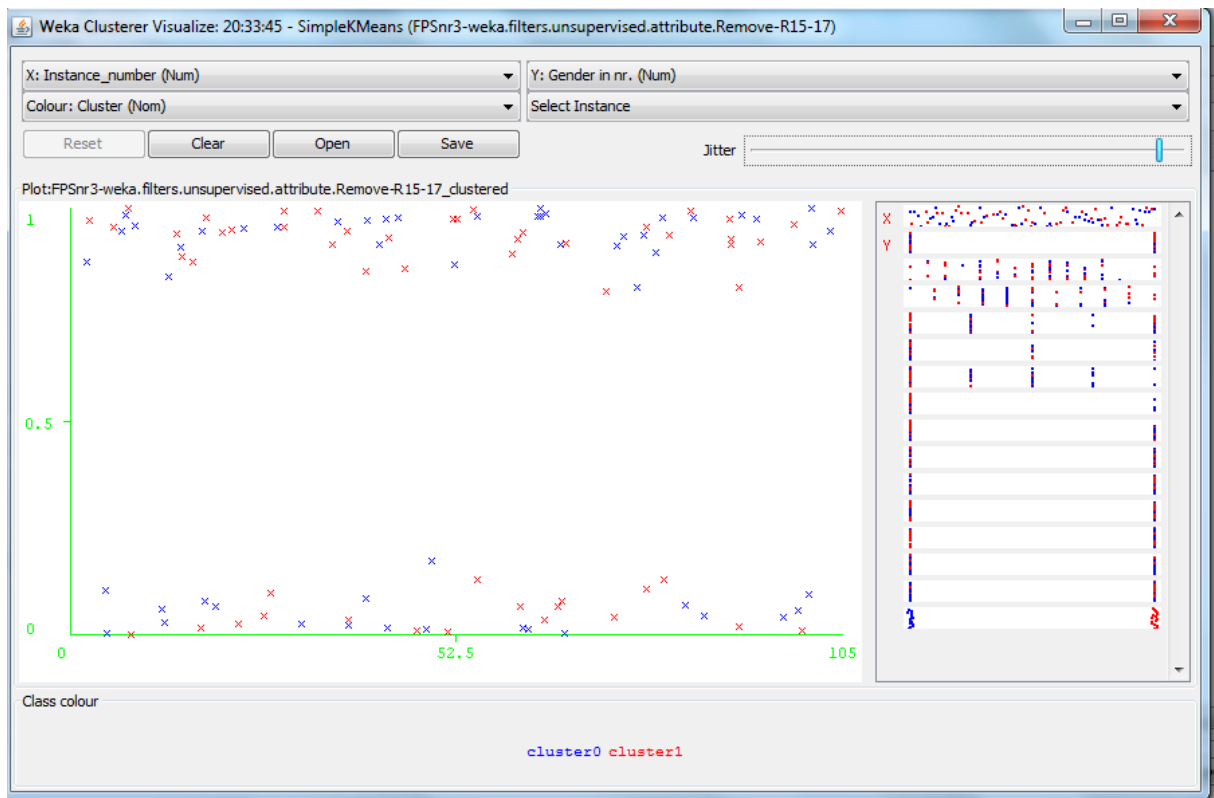


Figure 28 - Cluster visual inspection of fourth experiment

4.2.5 Selecting the best Clustering Model

The three criteria's we will put under consideration when choosing the best cluster model are Within cluster sum of squared error values, Number of iteration and the time which takes to build the model.

Within cluster sum of squared errors determines the tightness of cluster model, the lower gets it's value the better choice it becomes. It's used as a mechanism for assessing the goodness of the cluster model. Number of Iteration of the algorithm tells us how many loops it took to assign the displaced data items to the appropriate classes. So the lower gets the value of the iteration the preferred choice it becomes, since that tells us the convergence of the algorithm was pretty fast.

Experiment number	Within cluster sum of squared error values	Number of iteration	Time taken to build the model
I	210.79	7	0.04
II	207.58	5	0
III	208.11	9	0.01
IV	353.03	3	0

Table – Comparing the four clustering models

The above table shows us the output of the four clustering experiments. Based on the results from the above table, the second (2) experiments seems the best available option since it has the smallest value of within cluster sum of squared errors, comparatively one of the lowest numbers of iteration and the least time to build the model, in comparison with the other three experiments. In the figure below, we can see the visualization of all the cluster assignments of our best clustering model.

Furthermore, the knowledge acquired from the newly constructed cluster model is essential when it comes to splitting the participant’s data into Trusted and Not-Trusted.

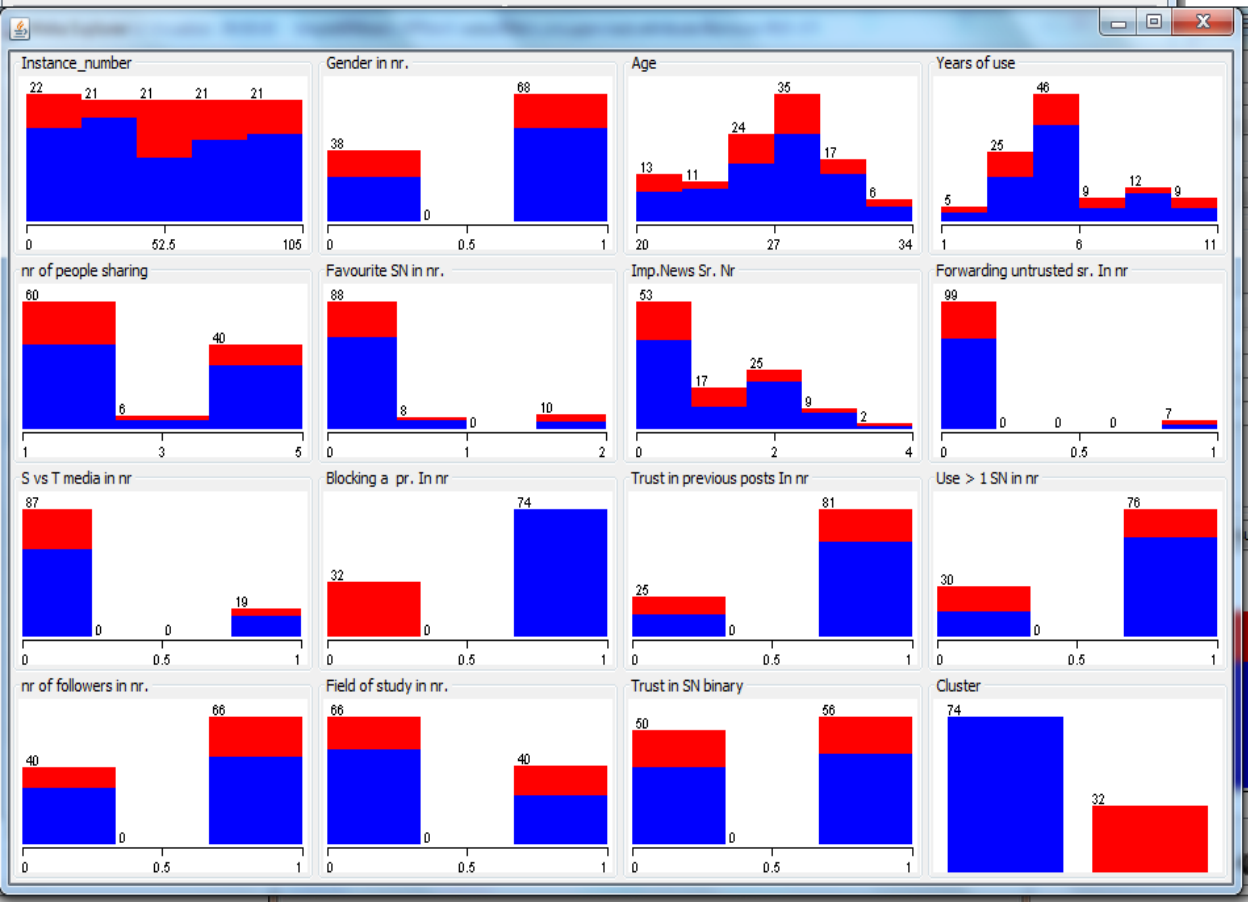


Figure 29 - Visualization of all the cluster assignments of experiment 2

4.3 Classification

As one of the main goals of this study is to predict trust using data mining techniques, a classification technique was adopted to develop a predictive model. The models were built with three different supervised machine learning algorithms i.e. Decision Tree Classification Algorithm, Bayesian Classifier and Neural Network using WEKA 3.6.11 machine learning software.

4.3.1 Experiment 1- J48 Decision tree

This experiment was performed to evaluate the performance of a J48 classifier decision tree in predicting trust to social media content. The decision tree algorithm was run on a full training set which contains 106 instances with 14 attributes. The amount of time which took to build the model is 0.04 seconds, and the model created a tree of size 37 with 19 leaves.

Type of Classification Model	Confusion Matrix		
	Distrust (Predicted)	Trust(Predicted)	Actual
J48 Unpruned	42	8	Distrust
	4	52	Trust

Table - The Confusion matrix result of J48 algorithm

The model also correctly classified 94 (88.68%) instances while 12 (11.32%) of the instances were also classified incorrectly. The comprehensive accuracy rate of the j48 model is profoundly successful, yet we should consider also the other factors like the TP Rate (Sensitivity), and TN Rate (Specificity) to evaluate the performance of the newly acquired model for each class.

This model has a TP Rate of 0.84, moreover the model has a tendency of identifying the negative occurrences as the FP Rate of the model is 0.071.

=== Summary ===

```

Correctly Classified Instances      94          88.6792 %
Incorrectly Classified Instances    12          11.3208 %
Kappa statistic                    0.7719
Mean absolute error                 0.1803
Root mean squared error             0.3002
Relative absolute error             36.1699 %
Root relative squared error         60.1431 %
Coverage of cases (0.95 level)     100 %
Mean rel. region size (0.95 level)  79.717 %
Total Number of Instances          106
  
```

=== Detailed Accuracy By Class ===

```

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0,840   0,071   0,913     0,840   0,875     0,774   0,937   0,929   Distrust
          0,929   0,160   0,867     0,929   0,897     0,774   0,937   0,926   Trust
Weighted Avg.   0,887   0,118   0,889     0,887   0,886     0,774   0,937   0,927
  
```

Figure 30 – Performance measures of J48

When it comes to Precision score of the model, around 91,3 % of participants were classified as associated to corresponding class Yes actually belong to class Yes, where as 86.7% of participants associated to class No actually belong to class No. Having an average precision of 88.9% this model turns out to be a highly successful model when it comes to labeling relevant values for individual class. Since this model has F-Measure value of 0.875 we can conclude that the Recall and the Precision of the model are to a large extent balanced.

Finally, as it is shown in results of this experiment the J48 decision tree algorithm is more than adequate in predicting trust for a social media content.

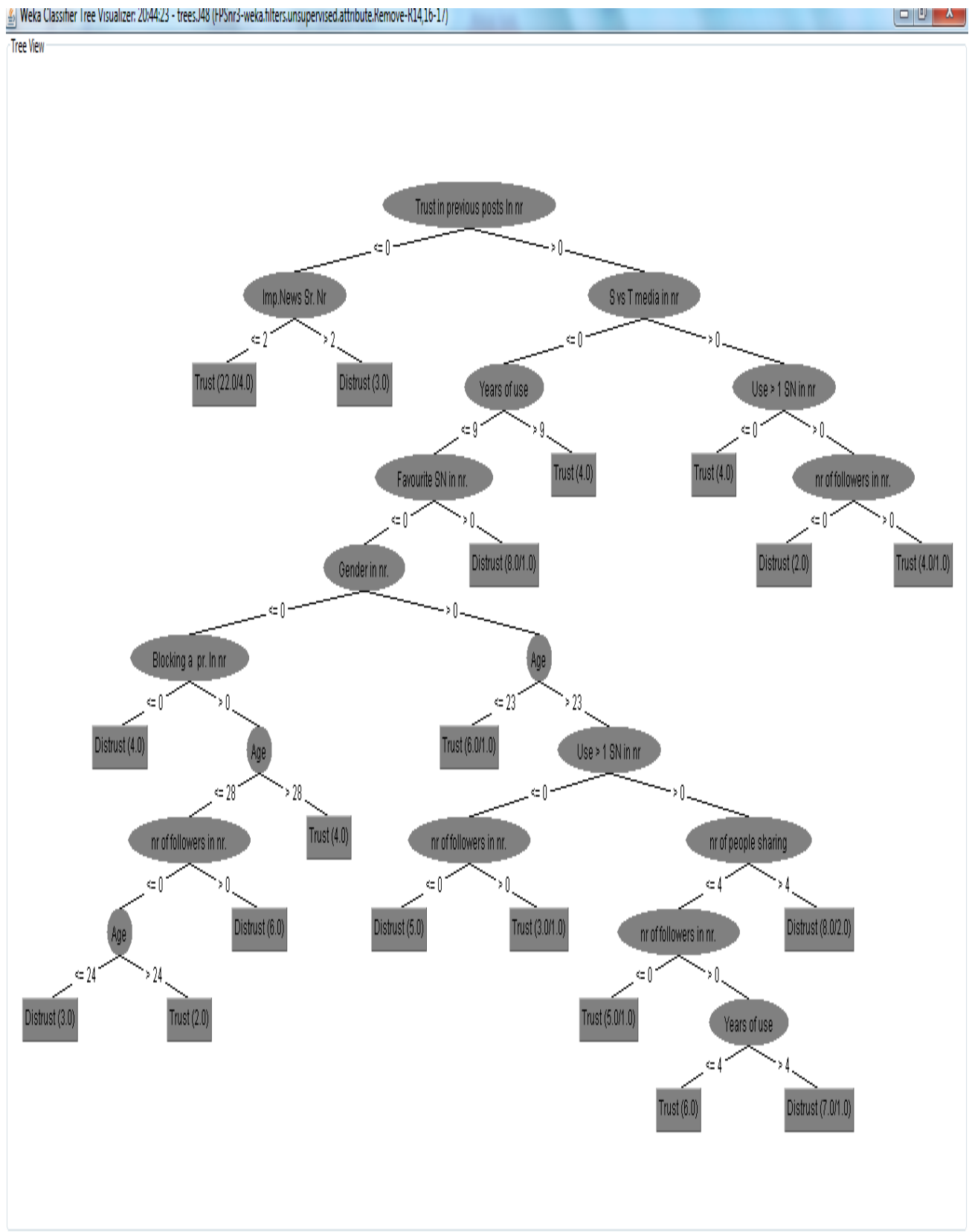


Figure 31 – Decision tree of the model

4.3.2 Experiment 2 – Naïve Bayes

The second experiment was performed to evaluate the performance of a Naive Bayes classifier in predicting trust to social media content. Naïve Bayes classifier was run on a full training set which contains 106 instances with 14 attributes. The amount of time which took to build the model is 0.02 seconds.

Type of Classification Model	Confusion Matrix		
	Distrust (Predicted)	Trust(Predicted)	Actual
Naive Bayes	32	18	Distrust
	14	42	Trust

Table - The Confusion matrix result of Naïve Bayes

The model also correctly classified 74 (69.81%) instances while 32 (30.19%) of the instances were also classified incorrectly. The comprehensive accuracy rate of the Naïve Bayes model is moderately successful, yet we should consider also the other factors like the TP Rate (Sensitivity), and TN Rate (Specificity) to evaluate the performance of the newly acquired model for each class.

This model has a TP Rate of 0.64; moreover the model has a tendency of identifying the negative occurrences as the FP Rate of the model is 0.25.

=== Summary ===

Correctly Classified Instances	74	69.8113 %
Incorrectly Classified Instances	32	30.1887 %
Kappa statistic	0.3917	
Mean absolute error	0.4171	
Root mean squared error	0.46	
Relative absolute error	83.6862 %	
Root relative squared error	92.1458 %	
Coverage of cases (0.95 level)	99.0566 %	
Mean rel. region size (0.95 level)	99.5283 %	
Total Number of Instances	106	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,640	0,250	0,696	0,640	0,667	0,393	0,733	0,696	Distrust
	0,750	0,360	0,700	0,750	0,724	0,393	0,733	0,739	Trust
Weighted Avg.	0,698	0,308	0,698	0,698	0,697	0,393	0,733	0,719	

Figure – Performance measures of Naïve bayes

When it comes to Precision score of the model, around 69,6 % of participants were classified as associated to corresponding class Yes actually belong to class Yes, where as 86.7% of participants associated to class No actually belong to class No. Having an average precision of 69.8% this model turns out to be a moderately successful model when it comes to labeling

relevant values for individual class. Since this model has F-Measure value of 0.667 we can conclude that the Recall and the Precision of the model are somehow balanced.

As it is shown in the confusion matrix, the model have 42 true positive,32 true negative, 18 false positive, and 14 false negative compounds.

The value of precision for trust compounds in this model is 0.7, which is quite ok. Moreover, the model has pretty good value of ROC Area for trust compounds, which is 0.733. Based on the results we can conclude that this Naïve Bayesian model could adequately be used for modeling trust to social media content.

4.3.3 Experiment 3- Neural Network

Our third experiment was done to evaluate the capability of Neural Network in predicting trust to social media content. Multilayer Perception which is one type of Neural Network was chosen to conduct this experiment. As in the previous experiments, this particular experiment has also 14 attributes and 106 instances. It took the algorithm 0.55 seconds to build the model.

Type of Classification Model	Confusion Matrix		
	Distrust (Predicted)	Trust(Predicted)	Actual
Multilayer perception	47	3	Distrust
	1	55	Trust

Table - The Confusion matrix result of Neural Network

The model also correctly classified 102 (96.23%) instances while 4 (3.77%) of the instances were also classified incorrectly. The comprehensive accuracy rate of the Multilayer Perception model is tremendously successful, yet we should consider also the other factors like the TP Rate (Sensitivity), and TN Rate (Specificity) to evaluate the performance of the newly acquired model for each class.

This model has a TP Rate of 0.94; moreover the model has a tendency of identifying the negative occurrences as the FP Rate of the model is 0.018.

=== Summary ===

```

Correctly Classified Instances      102          96.2264 %
Incorrectly Classified Instances     4           3.7736 %
Kappa statistic                     0.9241
Mean absolute error                 0.0627
Root mean squared error             0.184
Relative absolute error             12.5704 %
Root relative squared error         36.8581 %
Coverage of cases (0.95 level)     97.1698 %
Mean rel. region size (0.95 level) 58.4906 %
Total Number of Instances          106

```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,940	0,018	0,979	0,940	0,959	0,925	0,945	0,931	Distrust
	0,982	0,060	0,948	0,982	0,965	0,925	0,945	0,886	Trust
Weighted Avg.	0,962	0,040	0,963	0,962	0,962	0,925	0,945	0,907	

Figure – Performance measures of Naïve bayes

When it comes to Precision score of the model, around 97,9 % of participants were classified as associated to corresponding class Yes actually belong to class Yes, where as 86.7% of participants associated to class No actually belong to class No. Having an average precision of 96.3% this model turns out to be a moderately successful model when it comes to labeling relevant values for individual class. Since this model has F-Measure value of 0.959 we can conclude that the Recall and the Precision of the model are somehow balanced.

4.3.4 The final chosen rules by using “Type of Trust” (ToT) as a targeted class are as follows

RULE 1, IF Trust in previous posts In nr ≤ 0 AND Imp.News Sr. Nr ≤ 2 AND Use > 1 SN in nr > 0

Then ToT: Trust (14.0/1.0)

RULE 2, IF Gender in nr. ≤ 0 AND Use > 1 SN in nr > 0 AND nr of followers in nr. > 0

Then ToT: Distrust (10.0/1.0)

RULE 3, IF Favorite SN in nr. > 0 AND Years of use ≤ 7 AND Blocking a pr. In nr > 0

Then ToT: Distrust (8.0)

RULE 4, IF Favorite SN in nr. > 0 AND Trust in previous posts In nr > 0

Then ToT: Trust (5.0)

RULE 5, IF forwarding un trusted sr. In nr > 0

Then ToT: Distrust (6.0/2.0)

RULE 6, IF Field of study in nr. ≤ 0 AND Years of use ≤ 2

Then ToT: Distrust (3.0)

RULE 7, IF Field of study in nr. ≤ 0 AND Blocking a pr. In nr ≤ 0 AND Use > 1 SN in nr ≤ 0 AND nr of people sharing ≤ 4 AND Age ≤ 26

Then ToT: Distrust (3.0/1.0)

RULE 8, IF Field of study in nr. ≤ 0 AND Blocking a pr. In nr > 0 AND Age ≤ 30 AND Trust in previous posts In nr > 0 AND S vs T media in nr ≤ 0 AND nr of followers in nr. > 0

Then ToT: Trust (9.0/1.0)

RULE 9, IF Field of study in nr. ≤ 0 AND Blocking a pr. In nr ≤ 0

Then ToT: Trust (6.0)

RULE 10, IF Blocking a pr. In nr ≤ 0 AND Imp.News Sr. Nr ≤ 0

Then ToT: Distrust (4.0)

RULE 11, IF Blocking a pr. In nr > 0 AND nr of followers in nr. ≤ 0 AND Gender in nr. > 0 AND Field of study in nr. ≤ 0 AND Imp.News Sr. Nr ≤ 1 AND Age > 25

Then ToT: Distrust (4.0/1.0)

RULE 12, IF Blocking a pr. In nr > 0 AND nr of followers in nr. ≤ 0 AND Gender in nr. > 0

Then ToT: Trust (7.0)

RULE 13, IF Blocking a pr. In nr > 0 AND Use > 1 SN in nr > 0 AND Field of study in nr. <= 0 AND nr of people sharing > 1
Then ToT: Distrust (4.0)

RULE 14, IF Blocking a pr. In nr > 0 AND Field of study in nr. <= 0 AND Gender in nr. > 0
Then ToT: Distrust (5.0/2.0)

RULE 15, IF Field of study in nr. > 0 AND Blocking a pr. In nr > 0 AND Use > 1 SN in nr > 0 AND nr of people sharing <= 3
Then ToT: Distrust (5.0/1.0)

RULE 16, IF nr of followers in nr. > 0 AND SvsT media in nr <= 0 AND Years of use <= 8
Then ToT: Distrust (5.0/1.0)

4.3.5 Choosing the best classifier model

Subsequent to conducting the experiments the next step was comparing the models and choosing the best available model. The models were compared using different performance measures like time span, accuracy, TP Rate, FP Rate, F-Measure and ROC Area.

A brief summary of the performance of the three classification experiments is presented in the table below.

Type of Model	Accuracy	TP Rate	FP Rate	F-Measure	ROC Area	Time(Sec)
J48 un pruned with all attributes	88.68%	0.84	0.071	0.875	0.937	0.09
Naïve Bayes with all attributes	69.81%	0.64	0.25	0.667	0.733	0.01
Neural Network with all attributes	96.23%	0.94	0.018	0.959	0.945	0

Table- Comparison of the three classifier algorithms

Regarding the time which took to build the models, the Neural Network (Multilayer perception) classifier took the shortest time to build the models meanwhile, the experiment performed with Naïve Bayes scores the second best time, followed by J48 Decision tree classifier, which took the longest time of all the three algorithms.

When it comes to ROC Area, looking the area under the curve (AUC) to indicate the quality of separation, once again neural networks was the most accurate one, but also J48 Decision tree classifier outperforms Naïve Bayes classifiers to become the second best accuracy classifier.

Generally, Neural Network classifier outperformed the other two algorithms by achieving the fastest time and the best accuracy, TP-Rate, FP-Rate, and F-Measure and ROC Area values.

As a result, the model that is constructed with the Neural Network classification technique was taken as the final and binding classification model.

4.4 Regression Modeling

In this section, we have conducted and analyzed four kinds of regression models.

4.4.1 Linear regression

The result of the regression analysis is as follows:

Predictor	Coef	SE Coef	T	P
Constant	0,9473	0,4750	1,99	0,049
Gender in nr.	0,1174	0,1119	1,05	0,297
Age	-0,01814	0,01611	-1,13	0,263
Years of use	0,02211	0,02372	0,93	0,354
nr of people sharing	0,02071	0,03117	0,66	0,508
Favourite SN in nr.	-0,09652	0,08589	-1,12	0,264
Imp.News Sr. Nr	-0,02950	0,04483	-0,66	0,512
Forwarding untrusted sr. In nr	-0,2257	0,2223	-1,02	0,313
S vs T media in nr	0,1472	0,1440	1,02	0,309
Blocking a pr. In nr	0,0068	0,1198	0,06	0,955
Trust in previous posts In nr	-0,1769	0,1288	-1,37	0,173
Use > 1 SN in nr	0,0010	0,1214	0,01	0,994
nr of followers in nr.	0,0714	0,1055	0,68	0,500
Field of study in nr.	-0,1520	0,1086	-1,40	0,165

S = 0,501264 R-Sq = 12,5% R-Sq(adj) = 0,1%

In this case, the standard deviation of the error terms is 0,5. A 0,1% R-sq adj tells us that whenever there is an observation of a variation in the value of trust in social networks (dependent parameter), 12.5% of it is due to the model (or due to change in the independent parameters) and the remaining 87.5% is because of error or some other factor. This shows us our data doesn't fit well to the proposed linear model.

The regression equation is

$$\begin{aligned} \text{Trust in SN} = & 0,947 + 0,117 \text{ Gender in nr.} - 0,0181 \text{ Age} \\ & + 0,0221 \text{ Years of use} + 0,0207 \text{ nr of people sharing} \\ & - 0,0965 \text{ Favorite SN in nr.} - 0,0295 \text{ Imp.News Sr Nr} \\ & - 0,226 \text{ forwarding un trusted sr. In nr} \\ & + 0,147 \text{ SvsT media in nr} + 0,007 \text{ Blocking pr. In nr} \\ & - 0,177 \text{ Trust in previous posts In nr} \\ & + 0,001 \text{ Use>1 SN in nr} + 0,071 \text{ nr of followers nr.} \\ & - 0,152 \text{ Field of study in nr.} \end{aligned}$$

The equation represent a linear equation of the form, $Y = C + n_1X_1 + n_2X_2 + n_3X_3 + \dots$

This indicates that the resulting relation among the dependent and independent variables is linear.

The R-Sq, which is defined as the intensity of relationship is 12,5, indicates that

12.5% of the variations in Achievement is explained by the scores of the independent variables.

The P-values for the parameters used (Gender, Age, Years of use, nr. Of people sharing, Important news source , Forwarding un trusted info, Social Vs traditional media, Blocking a person, Trust in previous posts, Use >1 SN , number of followers and Field of study) are 0,297, 0,263, 0,354, 0,508, 0,264, 0,512, 0,313, 0,309, 0,955, 0,173, 0,994, 0,500 and 0,165 respectively.

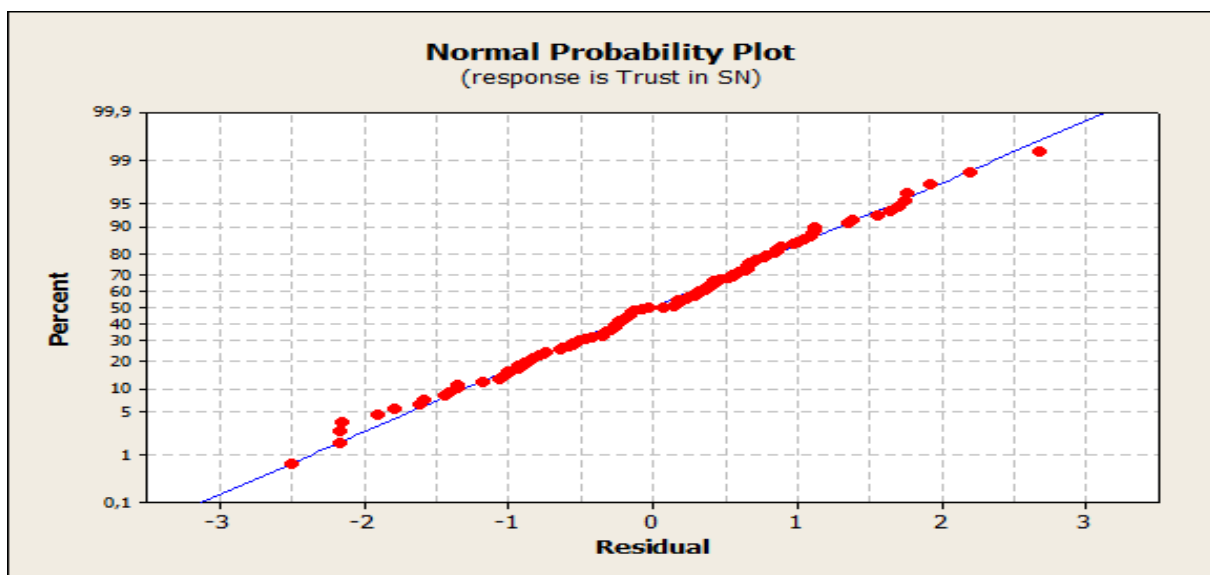
As we can see the p-values of the independent variables are not less than 0.05 indicating that there is no significant relationship in between independent variables and the dependent variable (Trust in SN).

T-stat value for all the independent parameters is less than 1.96 at 0.05 level of significance which indicates that there happens to be a no significant linear relationship in between the two parameters.

Analysis of Variance

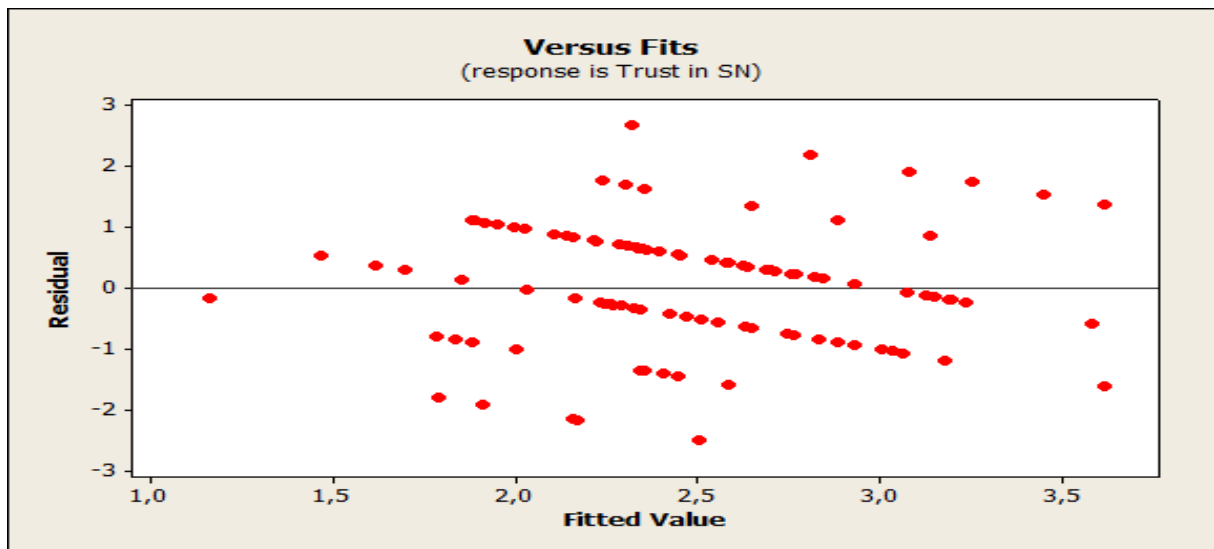
Source	DF	SS	MS	F	P
Regression	13	3,2987	0,2537	1,01	0,449
Residual Error	92	23,1164	0,2513		
Total	105	26,4151			

Figure 32 – Norm plot of Residuals for Trust in SN binary



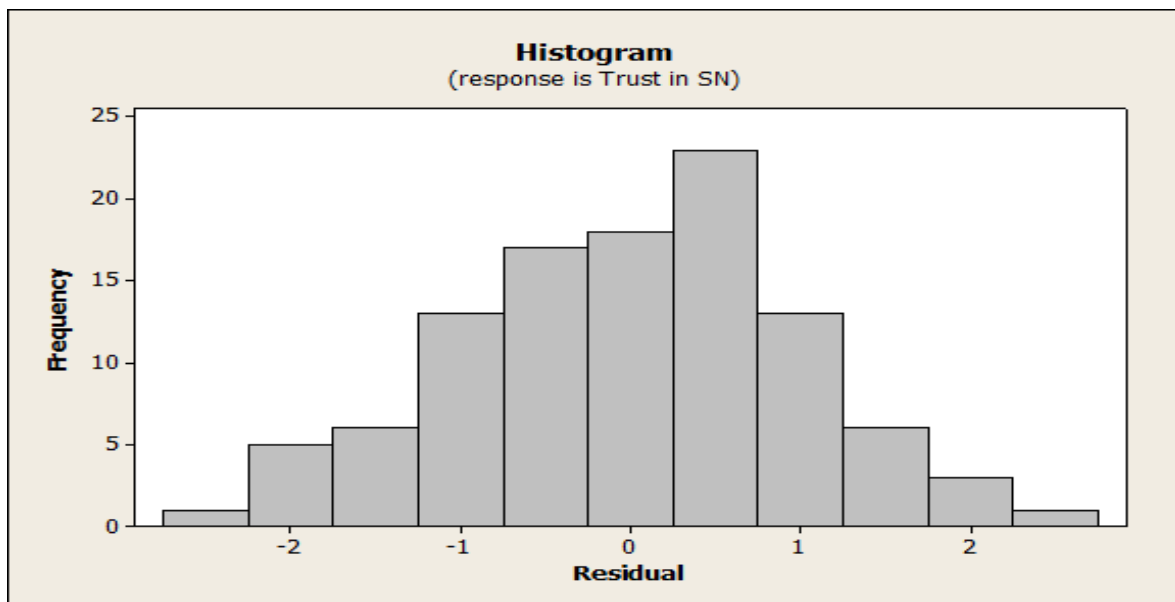
This graph checks the assumption of normality of error terms. We can clearly see that most of the red points are clustered around blue line, which indicates us the error terms are approximately normal. Thus our assumption of normality is valid.

Figure 33 - Residuals vs Fits for Trust in SN binary



Here, the graph plots the error terms against the fitted values. As we can see in the graph approximately half of them are above and the remaining half are below the zero line, which proves our assumption of the error terms having mean zero is valid.

Figure - Residual Histogram for Trust in SN binary



This graph again proves our normality assumption

4.4.2 Logistic regression - Minitab

In this experiment, we will conduct a logistic regression analysis in our data set by using Minitab and R. The reason for conducting the experiment by using both software's is that, even though Minitab gives a thorough analysis of binary logistic regression, it isn't possible to conduct Poisson and Negative binomial regression in Minitab. As a result, when it comes to choosing the best regression model, since the outputs of both software's are quite different it becomes preferable to do the analysis in both software's.

Table - Logistic Regression Table (Minitab)

Predictor	Coef	SE Coef	Z	P	Odds Ratio
Constant	2,17191	2,06805	1,05	0,294	
Gender in nr.	0,527662	0,476122	1,11	0,268	1,69
Age	-0,0859	0,0703115	-1,22	0,222	0,92
Years of use	0,102597	0,102593	1,00	0,317	1,11
nr of people sharing	0,0938015	0,134129	0,70	0,484	1,10
Favourite SN in nr.	-0,481409	0,389553	-1,24	0,217	0,62
Imp.News Sr. Nr	-0,146963	0,190598	-0,77	0,441	0,86
Forwarding untrusted	-1,04683	0,936633	-1,12	0,264	0,35
S vs T media in nr	0,742731	0,644067	1,15	0,249	2,10
Blocking a pr. In nr	0,0338852	0,520169	0,07	0,948	1,03
Trust in previous posts	-0,846473	0,575227	-1,47	0,141	0,43
Use > 1 SN in nr	0,0245806	0,538290	0,05	0,964	1,02
nr of followers in nr.	0,329272	0,446629	0,74	0,461	1,39
Field of study in nr.	-0,724665	0,476035	-1,52	0,128	0,48

Similar to any other regression analysis, we will start by checking the results of the p-values to determine if the predictor parameters have a significant relationship with the response parameter. Then, we will continue observing whether the coefficients have positive or negative relationship with the response parameter. As we can see from the above table, the parameters Gender, Years of use, nr. Of people sharing, Social Vs traditional media, Blocking a person, Use >1 SN, number of followers and Field of study have a positive relationship, while the remaining parameters have negative relationship towards the response parameter.

The odds ratio for Years of use is 1,11. If we assume that the other predictor variables to stay constant, for each one year increase in Years of use, the above model predicts an increase of 1.11 in the odds of the likelihood of the response being 1 to being a 0. In other words, whenever there is an increase of one year in Years of use, the response parameter is 1,11 times more likely to be a one than a zero.

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	107,315	92	0,131
Deviance	132,184	92	0,004
Hosmer-Lemeshow	5,691	8	0,682

Basically, there doesn't exist a model which has an exact fit. The thing we are interested in is that if the model is good enough for the purpose of analysis.

The logistic regression output indicates us that Deviance p-value of 0.004 give us significant evidence that our model fits well with our data. It means, our model reasonably describes the existing relationship in between the predicator and response parameters in the data set.

Figure 34 - Delta Chi-Square versus P

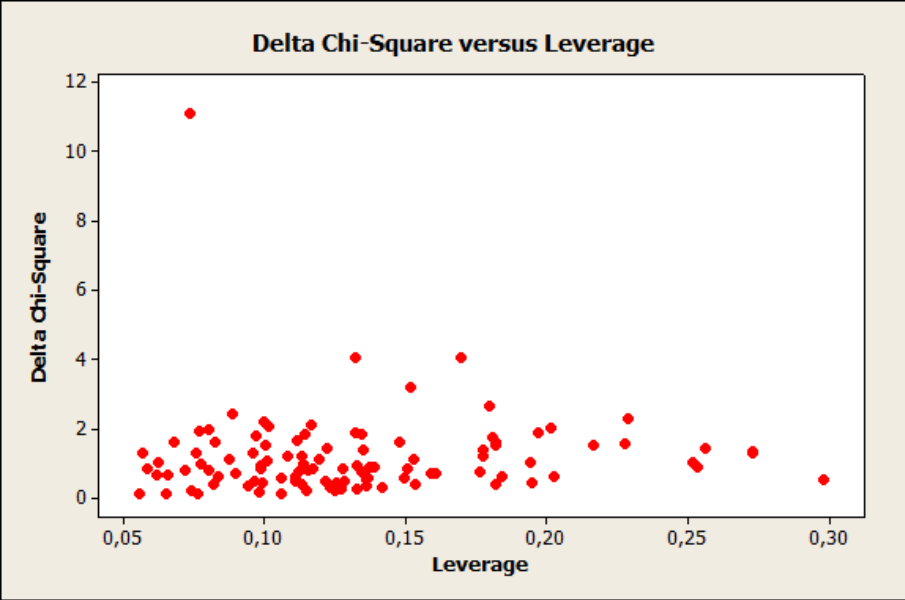


Figure 35- Delta Chi-Square versus Hi

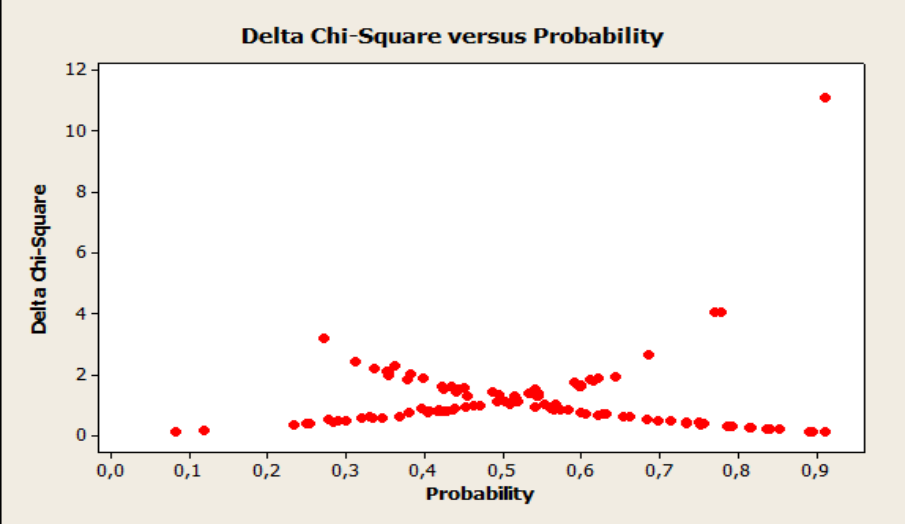
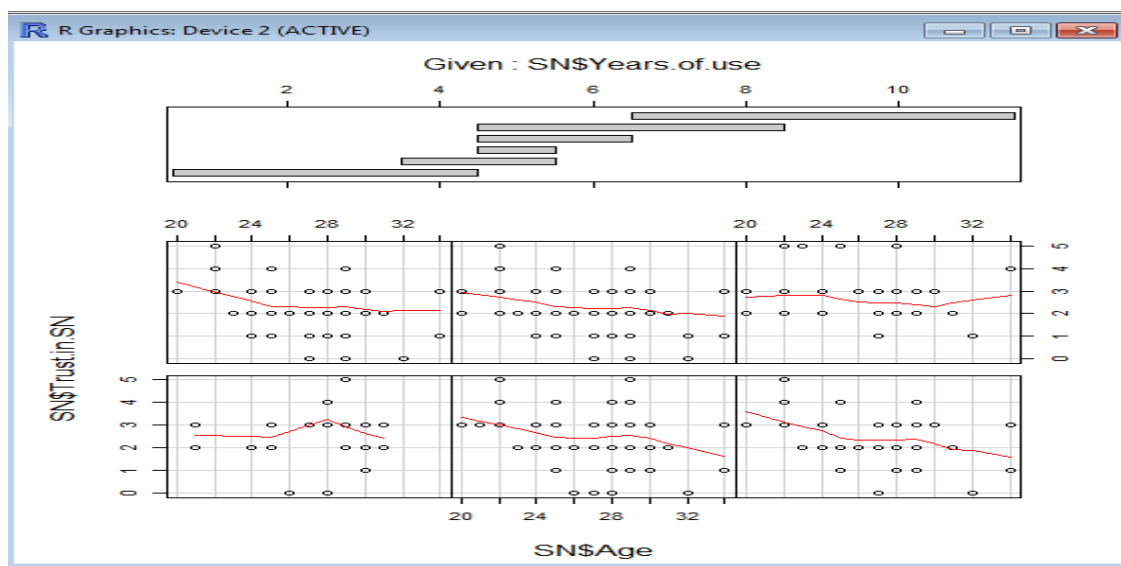


Figure 36 - The co-plot of trust against age and years of use in r



4.4.2.1 Logistic regression output in r

```
Call:
glm(formula = Trust.in.SN.binary ~ Age + Years.of.use + Gender.in.nr. +
     nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
     Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
     Blocking.a..pr..In.nr + Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
     nr.of.followers.in.nr. + Field.of.study.in.nr., family = binomial, data = SN)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2017  -1.0734   0.5227   1.0624   1.6166
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.17191    2.06798   1.050   0.294
Age              -0.08592    0.07031  -1.222   0.222
Years.of.use      0.10260    0.10259   1.000   0.317
Gender.in.nr.     0.52766    0.47611   1.108   0.268
nr.of.people.sharing 0.09380    0.13412   0.699   0.484
Favourite.SN.in.nr. -0.48141    0.38954  -1.236   0.217
Imp.News.Sr..Nr  -0.14696    0.19059  -0.771   0.441
Forwarding.untrusted.sr..In.nr -1.04683    0.93660  -1.118   0.264
S.vs.T.media.in.nr  0.74273    0.64404   1.153   0.249
Blocking.a..pr..In.nr  0.03389    0.52015   0.065   0.948
Trust.in.previous.posts.In.nr -0.84647    0.57520  -1.472   0.141
Use...1.SN.in.nr  0.02458    0.53827   0.046   0.964
nr.of.followers.in.nr.  0.32927    0.44662   0.737   0.461
Field.of.study.in.nr. -0.72466    0.47602  -1.522   0.128
```

(Dispersion parameter for binomial family taken to be 1)

The regression equation is

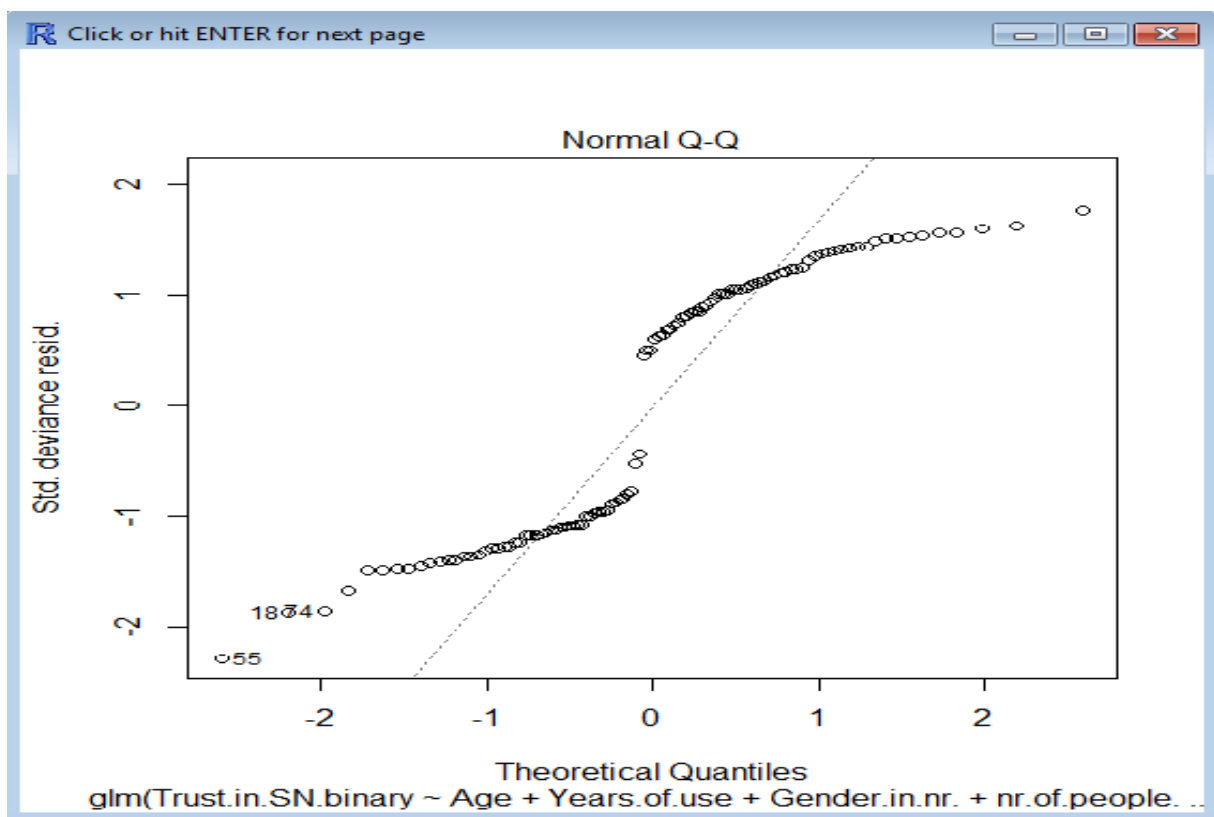
```
Trust in SN = 2.1719 + 0.52766 Gender in nr. - 0.08592 Age  
+ 0.1026 Years of use + 0.0938 nr of people sharing  
- 0.4814 Favorite SN in nr. - 0.1469 Imp.News Sr Nr  
- 1.0468 forwarding un trusted sr. In nr  
+ 0.743 SvsT media in nr + 0.0339 Blocking pr.In nr  
- 0.8465 Trust in previous posts In nr  
+ 0.0246 Use>1 SN in nr + 0.329 nr of followers nr.  
- 0.725 Field of study in nr.
```

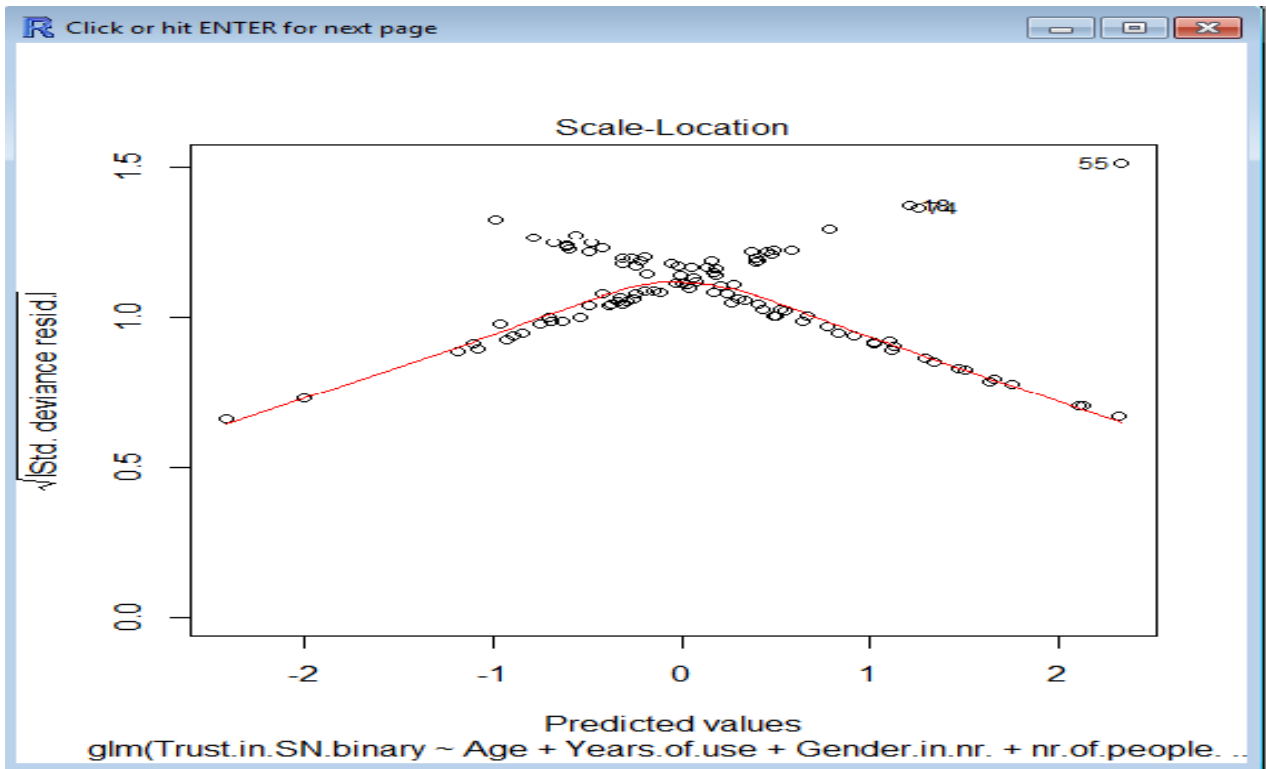
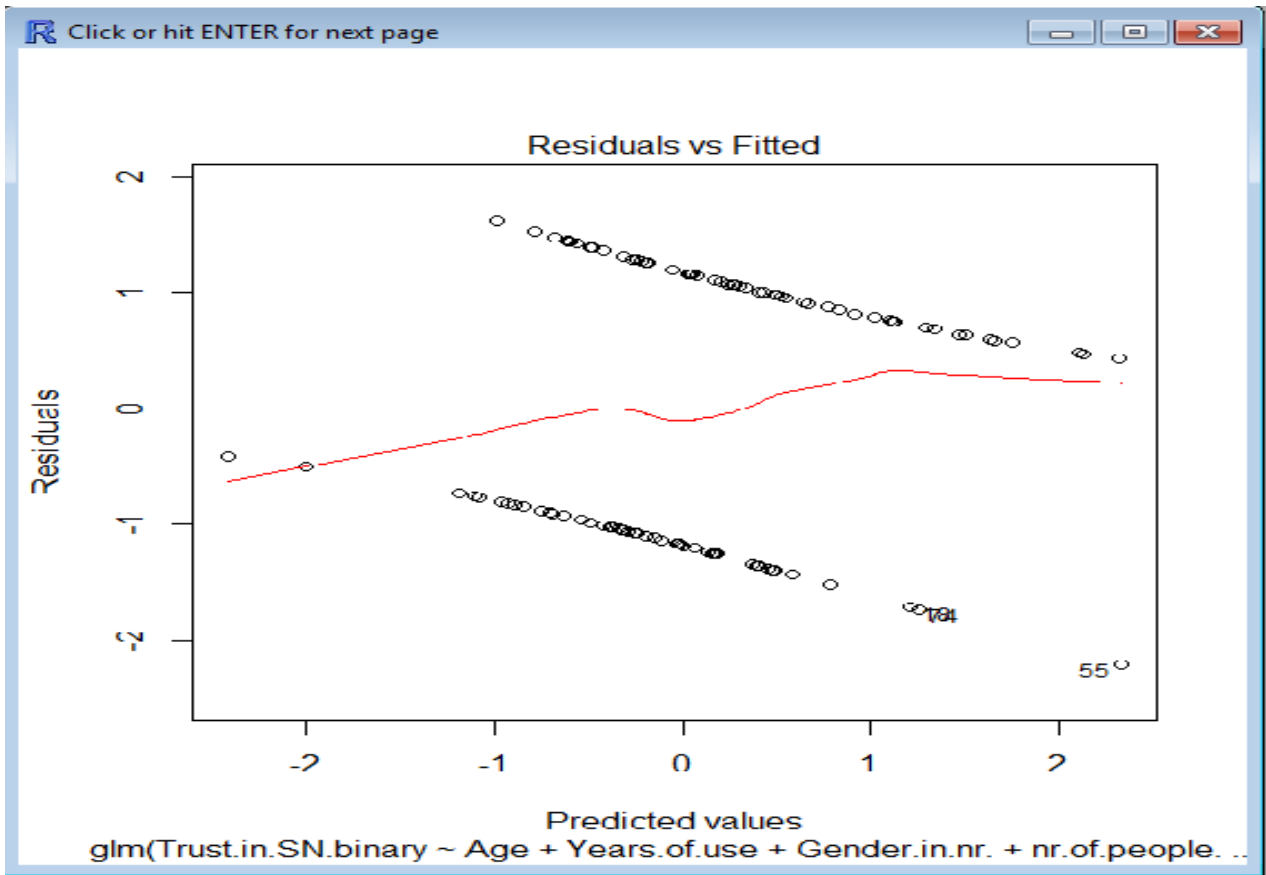
```
Null deviance: 146.61 on 105 degrees of freedom  
Residual deviance: 132.18 on 92 degrees of freedom  
AIC: 160.18
```

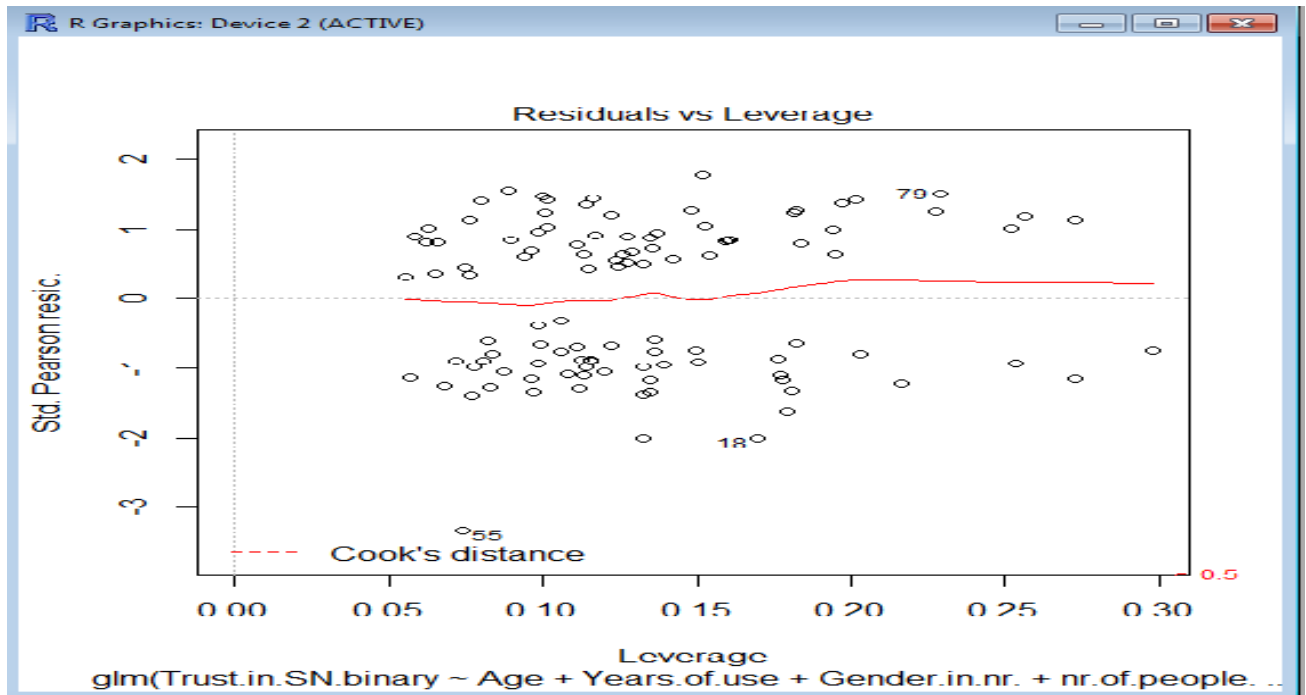
Number of Fisher Scoring iterations: 3

```
> 1-pchisq( 146.61, 105)  
[1] 0.004583104  
> 1-pchisq( 132.18, 92)  
[1] 0.003870083  
> 1-pchisq( 146.61 - 132.18, 105 - 92)  
[1] 0.344273
```

Figure 37 – plots of logistic regression







4.4.3 Poisson regression output in r

```
glm(formula = Trust.in.SN.binary ~ Age + Years.of.use + Gender.in.nr. +
     nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
     Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
     Blocking.a..pr..In.nr +
     Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
     nr.of.followers.in.nr. +
     Field.of.study.in.nr., family = Poisson, data = SN)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.48179	-0.94791	0.03353	0.54419	0.99993

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.07583	1.31637	0.058	0.954
Age	-0.03529	0.04413	-0.800	0.424
Years.of.use	0.04185	0.06517	0.642	0.521
Gender.in.nr.	0.24706	0.31831	0.776	0.438
nr.of.people.sharing	0.04104	0.08572	0.479	0.632
Favourite.SN.in.nr.	-0.22200	0.26478	-0.838	0.402
Imp.News.Sr..Nr	-0.04950	0.12665	-0.391	0.696
Forwarding.untrusted.sr..In.nr	-0.43945	0.66341	-0.662	0.508
S.vs.T.media.in.nr	0.25311	0.37062	0.683	0.495
Blocking.a..pr..In.nr	-0.01243	0.33799	-0.037	0.971
Trust.in.previous.posts.In.nr	-0.28005	0.33199	-0.844	0.399
Use...1.SN.in.nr	0.03091	0.33104	0.093	0.926
nr.of.followers.in.nr.	0.13560	0.29493	0.460	0.646
Field.of.study.in.nr.	-0.30331	0.31078	-0.976	0.329

(Dispersion parameter for poisson family taken to be 1)

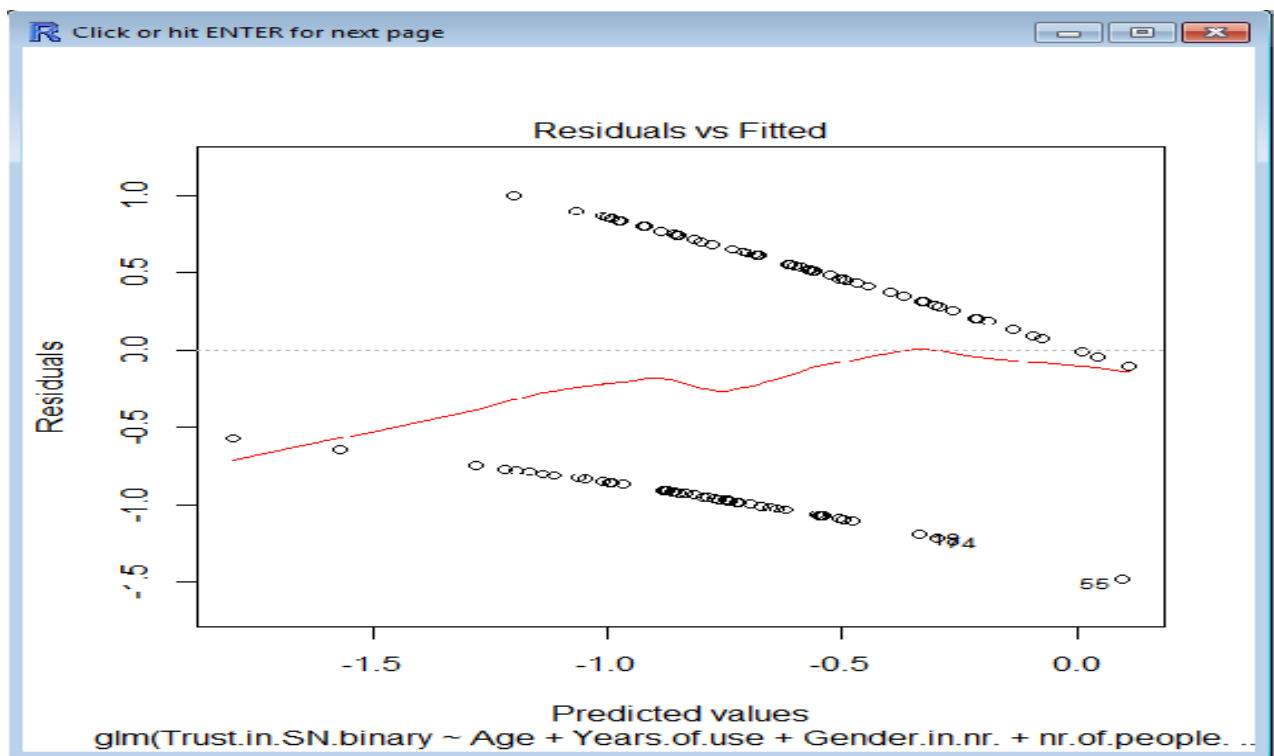
The regression equation is

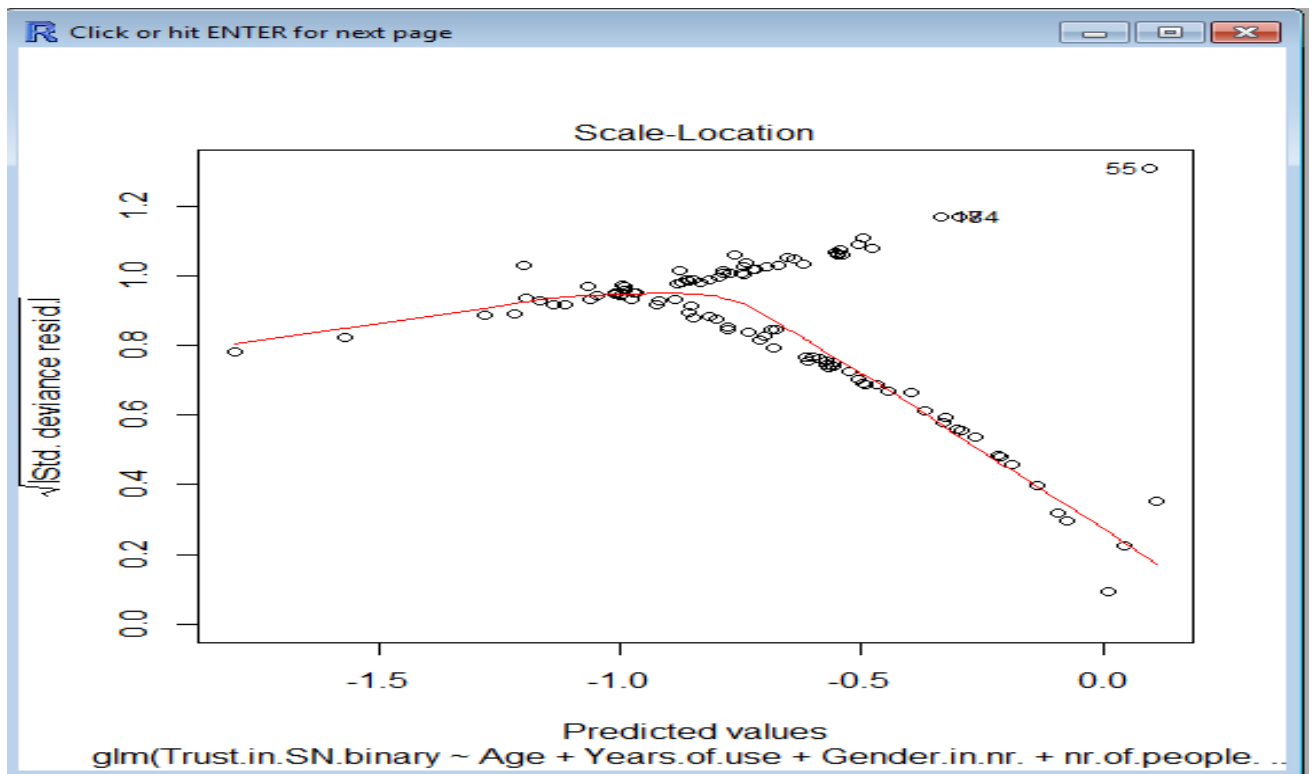
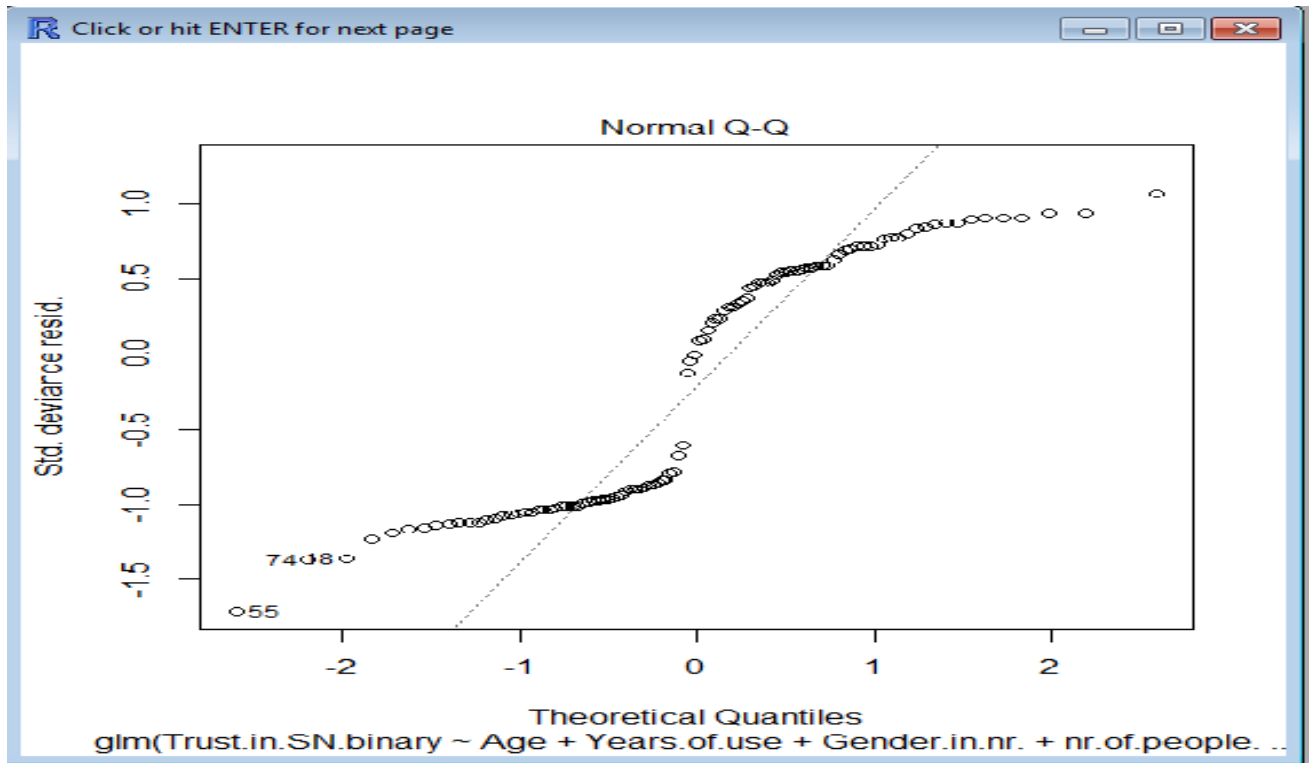
```
Trust in SN = 0.07583 + 0.24706 Gender in nr. - 0.03529 Age
              + 0.0419 Years of use + 0.0411 nr of people sharing
              - 0.222 Favorite SN in nr. - 0.0495 Imp.News Sr Nr
              - 0.43945 forwarding un trusted sr. In nr
              + 0.253 SvsT media in nr + 0.0124 Blocking pr.In nr
              - 0.28 Trust in previous posts In nr
              + 0.0309 Use>1 SN in nr + 0.136 nr of followers nr.
              - 0.3033 Field of study in nr.
```

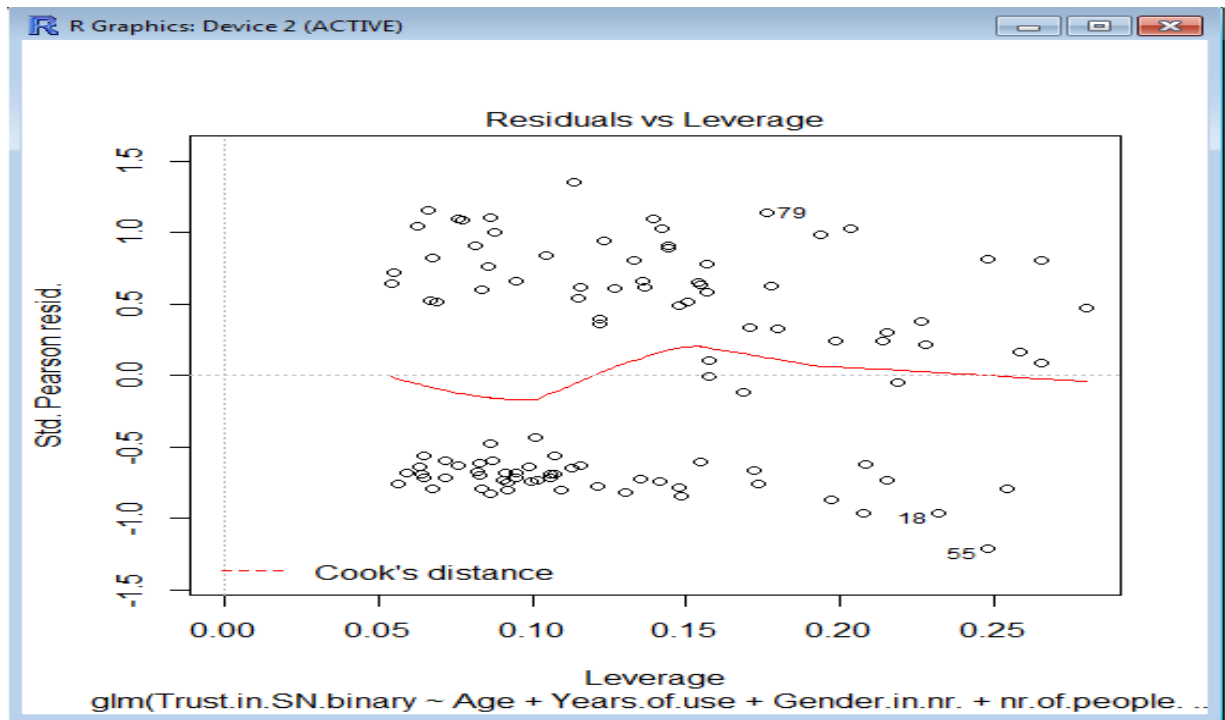
```
Null deviance: 71.466 on 105 degrees of freedom
Residual deviance: 65.328 on 92 degrees of freedom
AIC: 205.33
```

```
Number of Fisher Scoring iterations: 5
> 1-pchisq( 71.466, 105)
[1] 0.9949509
> 1-pchisq( 65.328, 92)
[1] 0.9840755
> 1-pchisq( 71.466 - 65.328, 105 - 92)
[1] 0.9409858
```

Figure38 – Plots of Poisson Regression







4.4.4 Negative binomial regression output in R

```
Call:
glm.nb(formula = Trust.in.SN.binary ~ Age + Years.of.use +
Gender.in.nr. +
nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
Blocking.a..pr..In.nr +
Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
nr.of.followers.in.nr. +
Field.of.study.in.nr., data = SN, init.theta = 19110.19472,
link = log)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.48178	-0.94791	0.03353	0.54417	0.99992

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.07582	1.31640	0.058	0.954
Age	-0.03529	0.04413	-0.800	0.424
Years.of.use	0.04185	0.06517	0.642	0.521
Gender.in.nr.	0.24706	0.31832	0.776	0.438

nr.of.people.sharing	0.04104	0.08573	0.479	0.632
Favourite.SN.in.nr.	-0.22200	0.26479	-0.838	0.402
Imp.News.Sr..Nr	-0.04950	0.12665	-0.391	0.696
Forwarding.untrusted.sr..In.nr	-0.43945	0.66342	-0.662	0.508
S.vs.T.media.in.nr	0.25311	0.37063	0.683	0.495
Blocking.a..pr..In.nr	-0.01242	0.33800	-0.037	0.971
Trust.in.previous.posts.In.nr	-0.28005	0.33199	-0.844	0.399
Use...1.SN.in.nr	0.03091	0.33105	0.093	0.926
nr.of.followers.in.nr.	0.13560	0.29494	0.460	0.646
Field.of.study.in.nr.	-0.30331	0.31078	-0.976	0.329

(Dispersion parameter for Negative Binomial(19110.19) family taken to be 1)

The regression equation is

Trust in SN = 0.07582 + 0.24706 Gender in nr. - 0.03529 Age
+ 0.0419 Years of use + 0.0411 nr of people sharing
- 0.222 Favorite SN in nr. - 0.0495 Imp.News Sr Nr
- 0.43945 forwarding un trusted sr. In nr
+ 0.253 SvsT media in nr + 0.0124 Blocking pr.In nr
- 0.28 Trust in previous posts In nr
+ 0.0309 Use>1 SN in nr + 0.136 nr of followers nr.
- 0.3033 Field of study in nr.

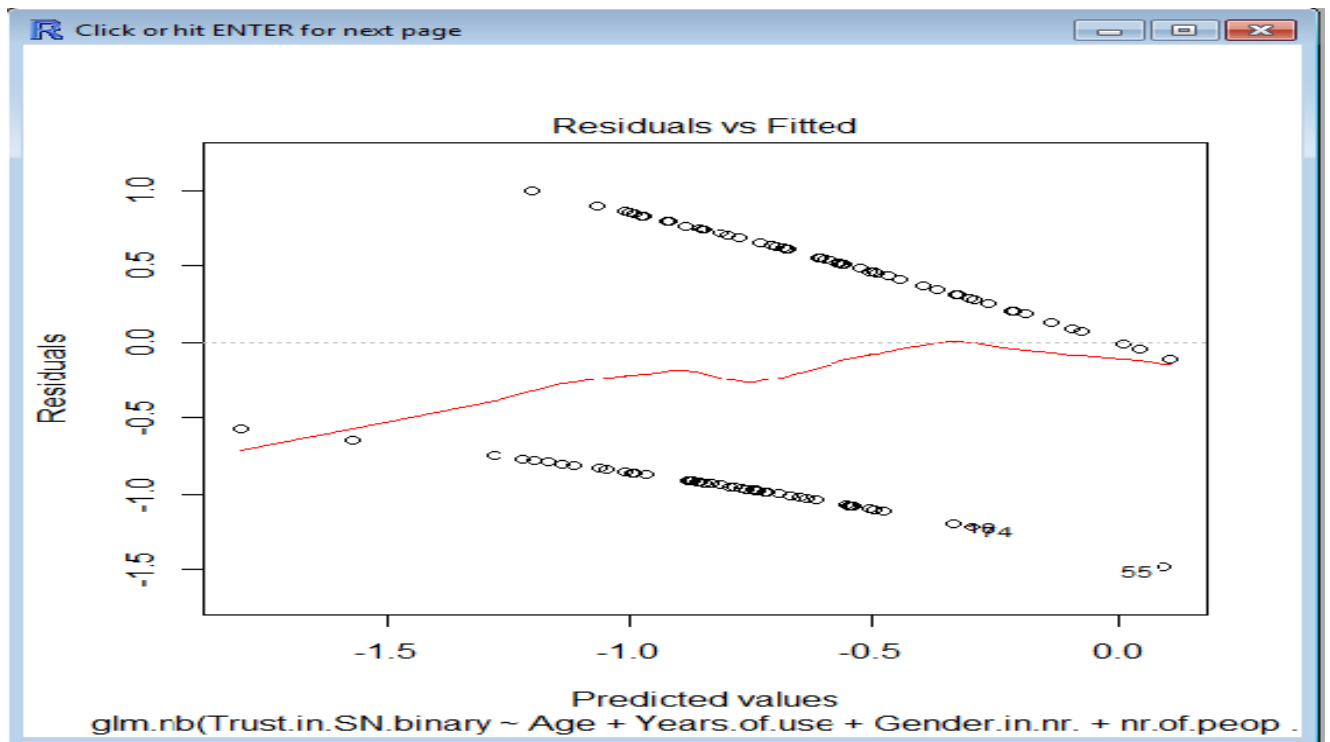
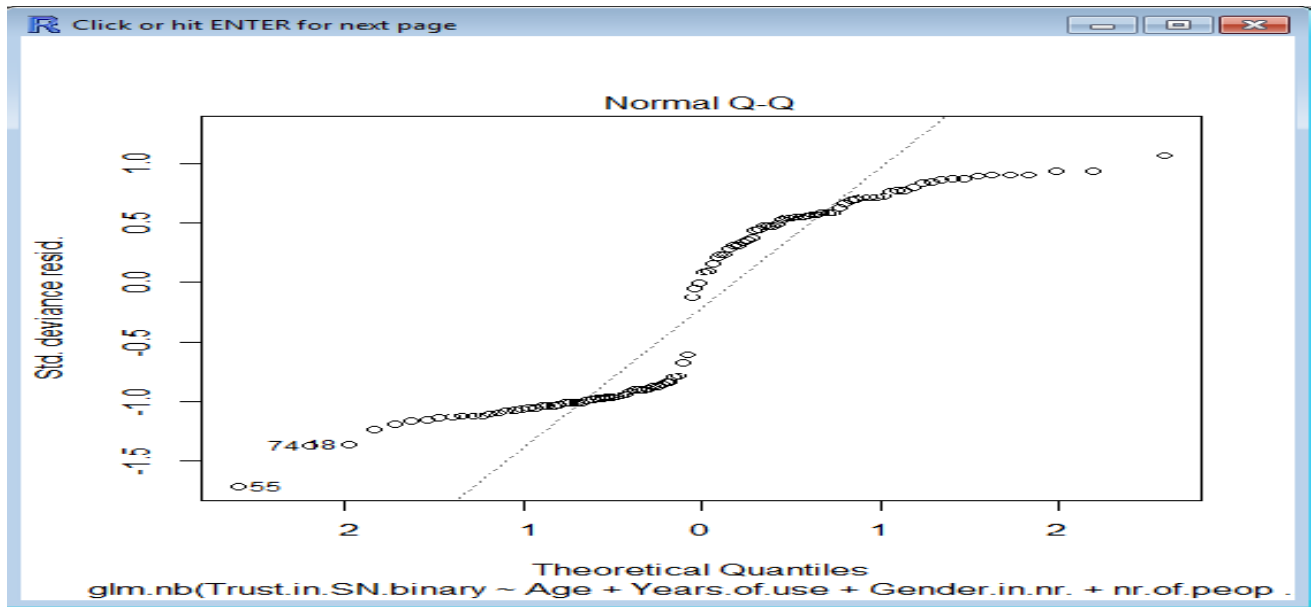
Null deviance: 71.464 on 105 degrees of freedom
Residual deviance: 65.327 on 92 degrees of freedom
AIC: 207.33

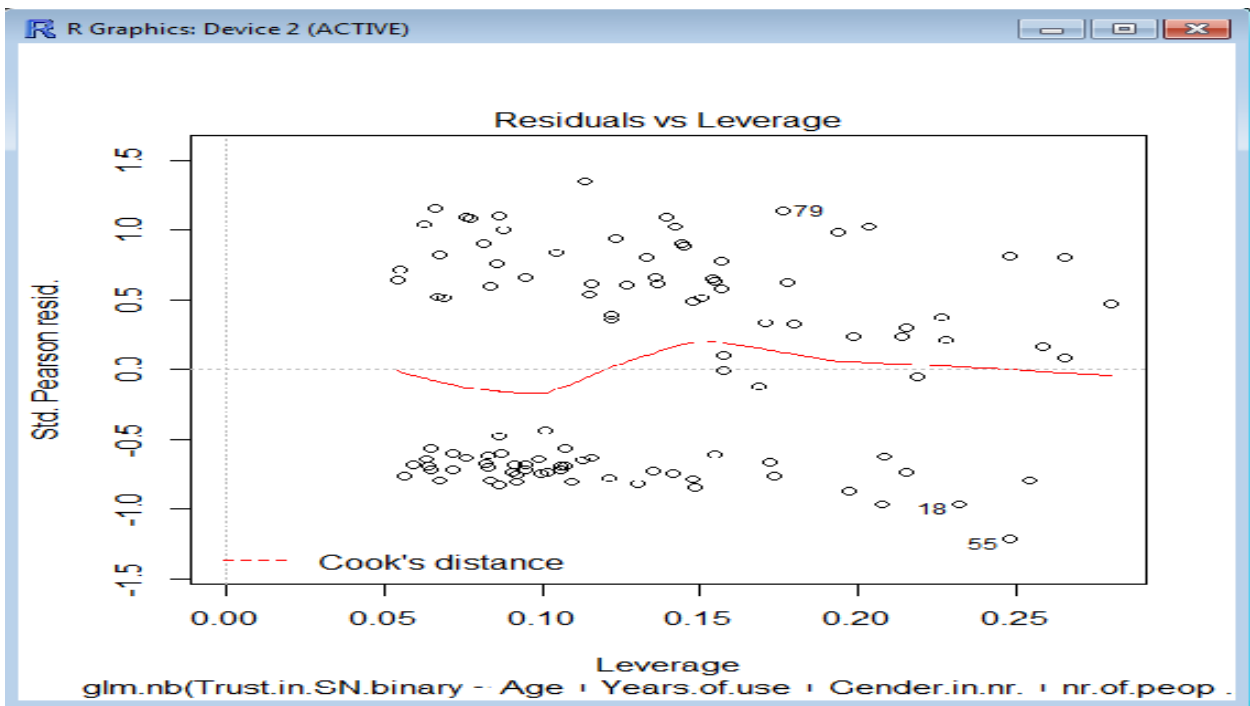
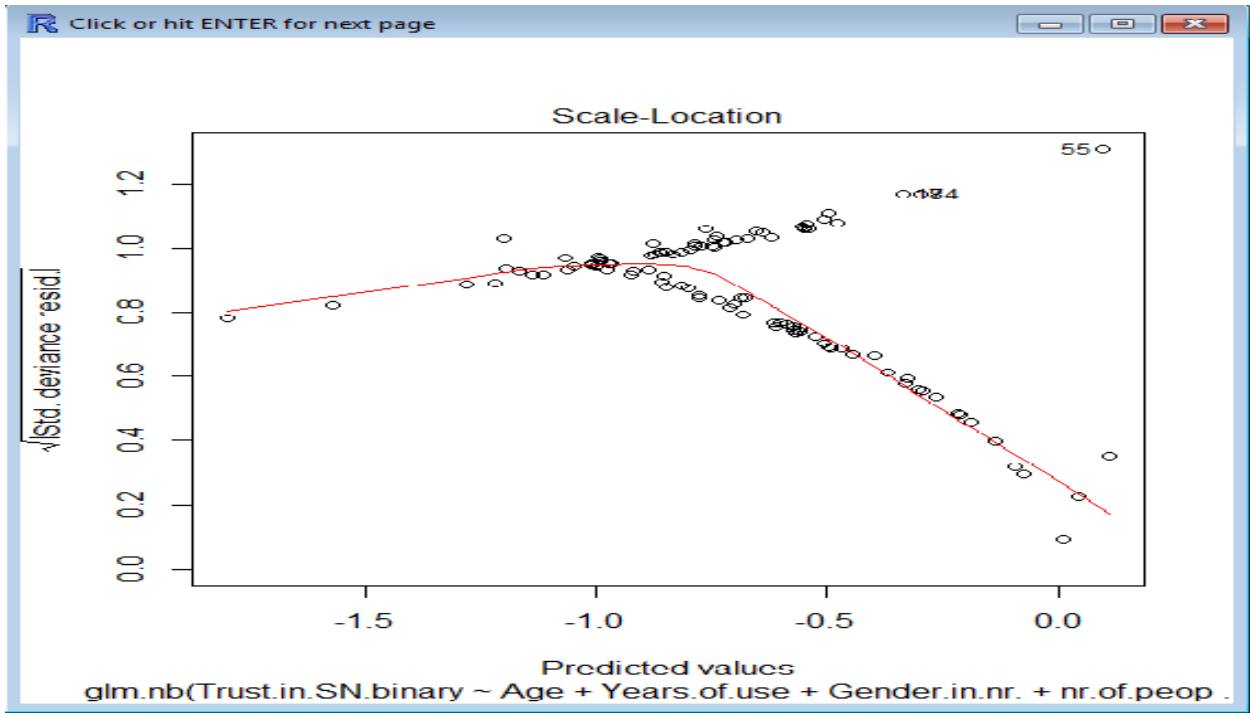
Number of Fisher Scoring iterations: 1

Theta: 19110
Std. Err.: 252042
Warning while fitting theta: iteration limit reached

2 x log-likelihood: -177.33
> 1-pchisq(71.464, 105)
[1] 0.9949535
> 1-pchisq(65.327, 92)
[1] 0.9840792
> 1-pchisq(71.464 - 65.327, 105 - 92)
[1] 0.9410243

Figure 39 – Plots of Negative binomial regression



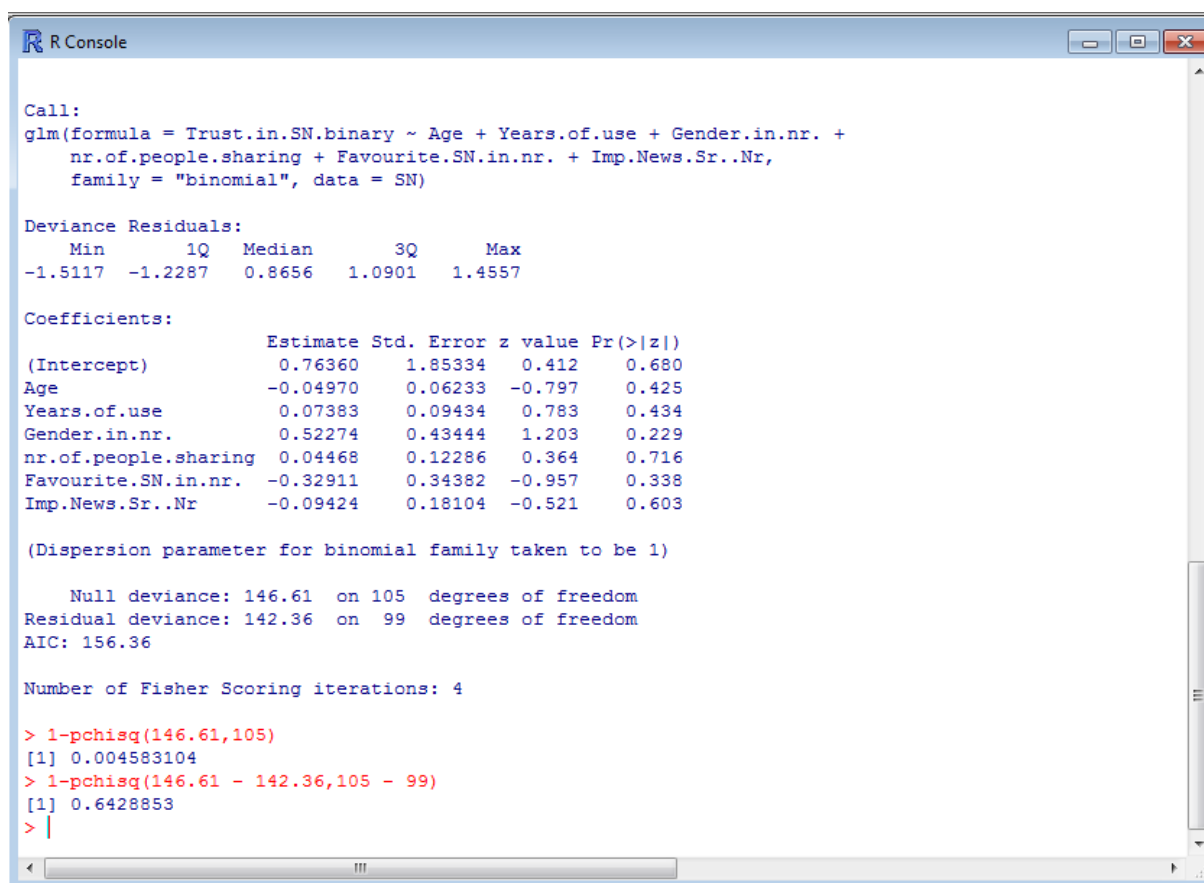


4.4.5 Comparison of the regression models

The research conducts experiments on linear regression, logistic regression, Poisson regression and negative binomial regression with in the survey data. After comparing the above mentioned models on the basis of AIC, log-likelihood and the two deviances(null and residual) the logistic regression model turns out to be the best alternative model.

The negative binomial regression reaches the iteration limit while fitting theta and gives a large value of theta as an output in addition to having high AIC value. Hence it is not a recommended model for this particular dataset.

Subsequent to selecting the best model available, the next step is removing the insignificant predictor variables from the model. For this purpose, I used the `glmulti()` function in R for automated model selection and model averaging.



```
R Console

Call:
glm(formula = Trust.in.SN.binary ~ Age + Years.of.use + Gender.in.nr. +
     nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr,
     family = "binomial", data = SN)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5117  -1.2287   0.8656   1.0901   1.4557

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.76360    1.85334   0.412  0.680
Age            -0.04970    0.06233  -0.797  0.425
Years.of.use    0.07383    0.09434   0.783  0.434
Gender.in.nr.   0.52274    0.43444   1.203  0.229
nr.of.people.sharing 0.04468    0.12286   0.364  0.716
Favourite.SN.in.nr. -0.32911    0.34382  -0.957  0.338
Imp.News.Sr..Nr -0.09424    0.18104  -0.521  0.603

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 146.61 on 105 degrees of freedom
Residual deviance: 142.36 on 99 degrees of freedom
AIC: 156.36

Number of Fisher Scoring iterations: 4

> 1-pchisq(146.61,105)
[1] 0.004583104
> 1-pchisq(146.61 - 142.36,105 - 99)
[1] 0.6428853
> |
```

Figure 40 – The final logistic regression model output

Model TYPE	AIC Value
Linear Regression	-
Logistic Regression	160.18
Poisson Regression	205.33
Negative binomial Regression	207.33

Table – Regression models with their respective AIC value

As it is shown in the figure above, the model has an improved values of AIC and p-values in comparison with the logistic regression model with all predictor parameters. The new model consists only 6 predictor parameters, unlike the first model which consists of all 14 parameters.

CHAPTER FIVE

5.1 Conclusion

This study showed that data mining techniques can be used efficiently to model and predict trust. The outcome of this study can be used to help people to make more consistent prediction of trust to social media content.

The data set used in this study was gathered from my own survey, which was prepared solely for the purpose of collecting data that can be used in this study. After the data was collected, it was preprocessed and prepared in a way suitable for the data mining tasks. Then the study was carried out in three sub phases, first the cluster modeling which then followed by classification modeling and finally regression modeling phase.

One of the main objective of this study was to conduct an experiment for observing how a person can decide on the trustworthiness of the information available in social media and to determine the significant factors that affect the trust to social media content. Some of the key findings from the study are listed below:

- The effect of engaging actively in social media on the overall trust is much weaker than originally predicted.
- Previous posts quality in social media is hugely influential when it comes to trust towards future posts of a particular user.
- The traditional media outlets are still more trusted than social media sites like Face book and twitter. Websites were found to be clear favorite as the most important news source by more than half of participants of the survey.
- Women tend to trust Social network sites as most important news source than men. Since 69 % of the participants who choose face book as their important news source were women. In addition, participants who have been members of social networks for more than five years tend to prefer social media outlets as their most important news source in comparison with those who have been members for less than five years.
- Even though the overwhelming majority of the participants have less trust in social media outlets in relation to traditional media outlets, they are still using social media outlets as their important source of news. Websites were found to be clear favorite as the most important news source by more than half of participants of the survey.

In this report we have also shown how different analysis can be drawn when we use focus group with members who have detailed technical know-how of the subject in contrast to the ordinary users which participated in the survey.

At last, based on the results of the conducted experiments the best alternative models for the three phases were chosen and the significant predictor parameters were found out.

- Logistic regression models turns out to be the best alternative regression model on the basis of AIC, log-likelihood and the two deviances(null and residual).
- The classifier model constructed with Neural Network classifier was selected as the most suitable classification model for this study.
- Years of use, important news source, Age, Favorite social network site, Gender and Number of people sharing in social media are significant attributes when it comes to determining trust to social media content.

For future research we will investigate different kinds of statistical methods to find more accurate measurement mechanism of trust and will make simulation experiments based on the findings. In this study we have done a survey of 108 people of age between 20 and 35, mainly consisting of university students, so our next step is to make a survey for a larger audience which consists of people from various demographic groups. In addition, we will try propose a model for recommendation, based on one of the popular kinds of social network sites.

References

- [1] (2010) Li Y. J, Dai Y. F. “Research on Trust Mechanism for Peerto-Peer Network”. Journal of Computers.
- [2] (2008) Wang J. C, Chiu C. C. “ Recommending trusted online auction sellers using social network analysis.” Expert Systems with Applications.
- [3] Nielsen. “U. S. Socail Media survey [C/OL] 2012”
- [4] Carrington P. J., Scott J., and Wasserman S., 2005. “Models and Methods in Social Network Analysis.” Cambridge University Press, New York, 2005.
- [5] Chang E., Thomson P., Dillon T., Hussain F., 2005. The Fuzzy and Dynamic Nature of Trust. S. Katsikas, J. López, G. Pernul (Eds.): TrustBus 2005, LNCS 3592, pp. 161-174, Springer – Verlag Berlin Heidelberg.
- [6] Liu Y., Yau S., Peng D., and Yin Y., 2008. “A Flexible Trust Model for Distributed Service Infrastructures.” In Proceedings of the 2008 11th IEEE Symposium on Object Oriented RealTime Distributed Computing, Orlando, USA, 108-115.
- [7] Rettinger A., Nickles M., and Tresp V., 2007. Learning Initial Trust among Interacting Agents. M. Klusch et al. (Eds.): CIA 2007, LNAI 4676, Springer – Verlag Berlin Heidelberg, pp. 313-327.
- [8] D. Boyd and N. Ellison. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, Jan 2008.
- [9] C. Charron, J. Favier, and C. Li. Social computing. How Networks Erode Institutional Power, Jan 2006.
- [10] Alan Mislove, Massilmiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. InProc.ofIMC,2007.
- [10] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User interactions in social networksand their implications. In Proc.ofEuroSys,2009.
- [11] Nichole, K. 2010, “4 ways to Measure Social Media and Its Impact on Your Brand”, Social Media Examiner, viewed 10th January 2012
- [12] Evans, D. 2008, “Social Media Marketing: An Hour a Day”, Wiley Publishing Inc., Indiana, United States.
- [13] Stelzner, M.A. 2012, “2012 Social Media Marketing Industry Report: How marketers are using Social Media to grow their businesses”, Social Media Examiner
- [14] Jennifer Golbeck. Computing and applying trust in web-based social networks. PhD

thesis, University of Maryland at College Park, 2005.

[15] Cai-Nicolas Ziegler and Georg Lausen. Propagation Models for Trust and Distrust in Social Networks. *Information Systems Frontiers*, 7(4-5):337–358, December 2005.

[16] John O’Donovan and Barry Smyth. Trust in recommender systems. *Proceedings of the 10th international conference on Intelligent user interfaces IUI 05*, 15:167, 2005.

[17] Hermida, A. (2010). From TV to Twitter: How ambient news became journalism. *Media/Culture Journal*, 13(2).

[18] Dwyer, C., Hiltz, S., Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. Paper presented at the Americas Conference on Information Systems, Colorado, CO. Retrieved June 20, 2011 from AIS Electronic Library (AISeL).

[19] http://changingminds.org/explanations/trust/what_is_trust.htm

[20] <http://www.definitions.net/definition/trust>

[21] Cai-Nicolas Ziegler and Georg Lausen (2005)
“PropagationModelsforTrustandDistrustinSocialNetworks

[22] Kim, Su-Yeon, Tae-Soo Jung, Eui-Ho Suh, and Hyun-Seok Hwang. 2006. Customer segmentation and strategy development based on customer lifetime value: a case study. *Expert Systems with Applications* 31: 101–107

[23] Koh, Chye Hian and Gerald Tan. n.d. Data mining applications in healthcare. *Journal of Healthcare Information Management* 2:64-72

[24] Kumar, Ela and Arun Solanki .2010. A combined mining approach and application in tax administration. *International Journal of Engineering and Technology* 2:38-44

[25] Kurgan, A. Lukasz, and Petr Musilek. 2006. A survey of knowledge discovery and data mining process models. United Kingdom: Cambridge University Press 21: 1-24

Mahler, J. Juliannem and Thomas Hennessey. 1996. Taking internal custom

[26] <http://www.texample.net/tikz/examples/neural-network/>

[27] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2):131–163, 1997.

[28] Jennifer Golbeck. Computing and applying trust in web-based social networks. PhD thesis, University of Maryland at College Park, 2005.

[29] Jennifer Golbeck. Personalizing applications through integration of inferred trust values in semantic web-based social networks. *W8: Semantic Network Analysis*, page 15, 2008.

Appendix

Questionnaire

1, Gender/ Kjønn

- Male/ Mann
- Female/ Kvinne

2, Are you part of a social network society? (Example - Face book, MySpace, tweeter) /

Er du en del av et nettbasert sosialt nettverk? (Eksempel - Facebook, MySpace, Twitter)

- Yes/ Ja
- No/ Nei

3, Is the number of people who commented or like a link which is shared in social media important for you when it comes to trusting the information. /

I vurderingen av å stole på innholdet i en link som er delt i et sosialt nettverk, er det viktig for deg å se hvor mange som har likt eller kommentert linken?

- Unimportant / Uviktig
- Less important/ Mindre viktig
- Neither / Verken eller
- Important/ Viktig
- Very Important/ Veldig viktig

4, Is Knowing the person who shared the information (it could be personally) important for you? /

Er det viktig for deg å kjenne personen som har delt informasjonen (det kan være personlig kjennskap)?

- Unimportant / Uviktig
- Less important/ Mindre viktig
- Neither / Verken eller
- Important/ Viktig
- Very Important/ Veldig viktig

5, Age / Alder

6, Do you think engaging actively in social media will make a person more trustworthy? /

du personens engasjerte aktivitet i et sosialt nettverk vil gjøre personen mer troverdig?

- Yes/ Ja
- No/ Nei

7, Do you use more than one social media networks./

Bruker du flere mediabaserte sosiale nettverk?

- Yes/ Ja
- No/ Nei

8, In your opinion, how important it is for a person to increase his trustworthiness by being actively engaged in more than one social media networks . /

Hvor viktig mener du det er at en person øker sin troverdighet gjennom å være engasjert og aktiv i mer enn ett sosialt nettverk?

- Unimportant / Uviktig
- Less important/ Mindre viktig
- Neither / Verken eller
- Important/ Viktig
- Very Important/ Veldig viktig

9, Is the number followers or friends the person sharing the information have influences your assessment of the credibility of the content. /

Har antall følgere og venner til personen stor påvirkningskraft for din vurdering av innholdets kredibilitet?

- Yes/ Ja
- No/ Nei

10, Does the trustworthiness of a person depends on the quality of the previous posts, comments and links he/she shares. /

Er personens troverdighet tilknyttet kvaliteten i de utleggene, kommentarene og linkene har eller hun har delt tidligere?

- Yes/ Ja
- No/ Nei

11, On average, how many people should share a content before you start trusting the information. /

Hvor mange folk burde, i gjennomsnitt, dele et innhold før du begynner å stole på informasjonen?

- 1-5
- 6-10
- 11- 15
- 16 – 20
- More than 20/ Mer enn 20

12, Do you think the information which is shared in social media is higher quality (trust worthy) than the traditional media outlets such as television, radio and newspapers? /

Tror du informasjonen som blir delt i sosiale nettverk er av høyere kvalitet og troverdighet enn hva som blir delt i de mer tradisjonelle medium (TV, radio, avis osv.)

- Yes/ Ja
- No/ Nei

**13, Which social media platform is your favorite? /
Hvilket mediabasert sosialt nettverk er din favoritt?**

- Twitter
- Face book
- MySpace
- Google+

14, Have you ever blocked or “unfriended “ people from your friends list because of the untrustworthiness of the information they share? /

Har du noen gang slettet noen fra din venneliste eller blokkert noen på grunn av deres mangel på troverdighet i innholdet de har delt?

- Yes/ Ja
- No/ Nei

**15, Which of the following is your most important news source? /
Hvilket medium er din viktigste nyhetskilde?**

- TV
- News paper / Avis
- Tweeter
- Face book
- Websites/ Nettsider
- Other / Andre

16, How much trust do you have in social media as a source of news? In a scale of 0 to 5 (5 if you fully trust them and 0 if you don't trust them at all). /

**Hvor mye stoler du på et sosialt medium som en nyhetskilde (på en skala fra 0-5)?
(5 om du stoler helt på det, og 0 om du ikke stoler på det i det hele tatt)**

- 0
- 1
- 2
- 3
- 4
- 5

17, How long have you been using social sites? (Example- 3 years) /

Hvor lenge har du brukt sosiale medium/nettverk (f.eks. 3 år)

18, What is your field of Study? / Hva studerer du?

19, Do you forward/share any content that you do not fully trust? / Deler eller videregiver du innhold som du ikke helt stoler på?

**20, Which of the following do you need to trust to a social media content? (you can select multiple) Please also order these criteria from the most important to the least. /
Hvilket av disse følgende punkt trenger du for å stole på innholdet i et sosialt medium?
Du kan velge flere alternativ, og vær vennlig og skriv kriteriene i rett rekkefølge, fra**

mest viktig til minst viktig.

- The source is known and well reputed by you / Kilden er velkjent og annerkjent av deg
- High number times the content is liked, shared and forwarded / Innholdet har blitt likt, delt og videresendt mange ganger
- Verified by conventional media/ Verifisert av konvensjonell media
- Verified by friends and colleagues / Verifisert av venner og kolleger
- Common sense or your intuition / Sunn fornuft/din intuisjon

21, Do you have any other criteria that you need to trust to a social media content? / du andre kriterium til grunn for å kunne stole på innholdet i et sosialt medium?

22, Which of the following make you NOT trust to social media content? (You can select multiple) Please also order these criteria from the most important to the least. / Hvilket av de følgende alternativene får deg til å IKKE stole på innholdet i et sosialt medium? Du kan velge flere alternativ

- Denial by the government or a governmental organization / Fornektelse fra staten eller statlige organisasjoner
- Denial by a trusted nongovernmental organization/ Fornektelse fra en troverdig ikke-statlig organisasjon
- Denial by the subject of the content/ Fornektelse på grunn av innholdets tema
- Number of denying social media content/ Antall som fornekter innholdet i et sosialt medium
- Inconsistent social media content/ Inkonsistent innhold i ett sosialt medium
- Inconsistent conventional media content/ Inkonsistent konvensjonell innhold i media
- Bad reputation of the source/ Dårlig rykte om kilden
- Common sense/your intuition / Sunn fornuft/din intuisjon

23, Do you have any other criteria that makes you NOT trust to a social media content?/ Har du noen andre kriterium som får deg til å IKKE stole på et innhold i et sosialt medium?

Appendix 1 J48 Classifier output in WEKA

==== Run information ====

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: FPSnr3-weka.filters.unsupervised.attribute.Remove-R14,16-17
Instances: 106
Attributes: 14
 Gender in nr.
 Age
 Years of use
 nr of people sharing
 Favourite SN in nr.
 Imp.News Sr. Nr
 Forwarding untrusted sr. In nr
 S vs T media in nr
 Blocking a pr. In nr
 Trust in previous posts In nr
 Use > 1 SN in nr
 nr of followers in nr.
 Field of study in nr.
 Trust in nominal

Test mode: evaluate on training data

==== Classifier model (full training set) ====

J48 pruned tree

```
Trust in previous posts In nr <= 0
| Imp.News Sr. Nr <= 2: Trust (22.0/4.0)
| Imp.News Sr. Nr > 2: Distrust (3.0)
Trust in previous posts In nr > 0
| S vs T media in nr <= 0
| | Years of use <= 9
| | | Favourite SN in nr. <= 0
| | | | Gender in nr. <= 0
| | | | | Blocking a pr. In nr <= 0: Distrust (4.0)
| | | | | Blocking a pr. In nr > 0
| | | | | Age <= 28
| | | | | | nr of followers in nr. <= 0
| | | | | | Age <= 24: Distrust (3.0)
| | | | | | Age > 24: Trust (2.0)
```

```

| | | | | | | nr of followers in nr. > 0: Distrust (6.0)
| | | | | | | Age > 28: Trust (4.0)
| | | | | | | Gender in nr. > 0
| | | | | | | Age <= 23: Trust (6.0/1.0)
| | | | | | | Age > 23
| | | | | | | Use > 1 SN in nr <= 0
| | | | | | | nr of followers in nr. <= 0: Distrust (5.0)
| | | | | | | nr of followers in nr. > 0: Trust (3.0/1.0)
| | | | | | | Use > 1 SN in nr > 0
| | | | | | | nr of people sharing <= 4
| | | | | | | nr of followers in nr. <= 0: Trust (5.0/1.0)
| | | | | | | nr of followers in nr. > 0
| | | | | | | Years of use <= 4: Trust (6.0)
| | | | | | | Years of use > 4: Distrust (7.0/1.0)
| | | | | | | nr of people sharing > 4: Distrust (8.0/2.0)
| | | | | | | Favourite SN in nr. > 0: Distrust (8.0/1.0)
| | | | | | | Years of use > 9: Trust (4.0)
| | | | | | | S vs T media in nr > 0
| | | | | | | Use > 1 SN in nr <= 0: Trust (4.0)
| | | | | | | Use > 1 SN in nr > 0
| | | | | | | nr of followers in nr. <= 0: Distrust (2.0)
| | | | | | | nr of followers in nr. > 0: Trust (4.0/1.0)

```

Number of Leaves : 19

Size of the tree : 37

Time taken to build model: 0.16 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.09 seconds

=== Summary ===

Correctly Classified Instances	94	88.6792 %
Incorrectly Classified Instances	12	11.3208 %
Kappa statistic	0.7719	
Mean absolute error	0.1803	
Root mean squared error	0.3002	
Relative absolute error	36.1699 %	
Root relative squared error	60.1431 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	79.717 %	
Total Number of Instances	106	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Class								
	0,840	0,071	0,913	0,840	0,875	0,774	0,937	Distrust
	0,929	0,160	0,867	0,929	0,897	0,774	0,937	Trust
Weighted Avg.	0,887	0,118	0,889	0,887	0,886	0,774	0,937	0,927

==== Confusion Matrix ====

```

a b <-- classified as
42 8 | a = Distrust
4 52 | b = Trust

```

Appendix 2 Naive Bayes Classifier output in WEKA

==== Run information ====

```

Scheme:   weka.classifiers.bayes.NaiveBayes
Relation: FPSnr3-weka.filters.unsupervised.attribute.Remove-R14,16-17
Instances: 106
Attributes: 14
  Gender in nr.
  Age
  Years of use
  nr of people sharing
  Favourite SN in nr.
  Imp.News Sr. Nr
  Forwarding untrusted sr. In nr
  S vs T media in nr
  Blocking a pr. In nr
  Trust in previous posts In nr
  Use > 1 SN in nr
  nr of followers in nr.
  Field of study in nr.
  Trust in nominal

```

Test mode: evaluate on training data

==== Classifier model (full training set) ====

Naive Bayes Classifier

Attribute	Class	
	Distrust	Trust
	(0.47)	(0.53)

```

=====
Gender in nr.
  mean          0.58  0.6964
  std. dev.     0.4936 0.4598
  weight sum      50   56
  precision      1    1

```

Age		
mean	27.3969	27.0385
std. dev.	3.1667	3.5641
weight sum	50	56
precision	1.0769	1.0769

Years of use		
mean	5.52	5.8214
std. dev.	1.9208	2.4061
weight sum	50	56
precision	1	1

nr of people sharing		
mean	2.8	2.8571
std. dev.	1.7205	1.6194
weight sum	50	56
precision	1	1

Favourite SN in nr.		
mean	0.34	0.1964
std. dev.	0.6815	0.5484
weight sum	50	56
precision	1	1

Imp.News Sr. Nr		
mean	1.04	0.8929
std. dev.	1.1993	1.0295
weight sum	50	56
precision	1	1

Forwarding untrusted sr. In nr		
mean	0.08	0.0536
std. dev.	0.2713	0.2252
weight sum	50	56
precision	1	1

S vs T media in nr		
mean	0.12	0.2321
std. dev.	0.325	0.4222
weight sum	50	56
precision	1	1

Blocking a pr. In nr		
mean	0.68	0.7143
std. dev.	0.4665	0.4518
weight sum	50	56
precision	1	1

Trust in previous posts In nr		
mean	0.86	0.6786

std. dev.	0.347	0.467
weight sum	50	56
precision	1	1

Use > 1 SN in nr

mean	0.74	0.6964
std. dev.	0.4386	0.4598
weight sum	50	56
precision	1	1

nr of followers in nr.

mean	0.58	0.6607
std. dev.	0.4936	0.4735
weight sum	50	56
precision	1	1

Field of study in nr.

mean	0.44	0.3214
std. dev.	0.4964	0.467
weight sum	50	56
precision	1	1

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances	74	69.8113 %
Incorrectly Classified Instances	32	30.1887 %
Kappa statistic	0.3917	
Mean absolute error	0.4171	
Root mean squared error	0.46	
Relative absolute error	83.6862 %	
Root relative squared error	92.1458 %	
Coverage of cases (0.95 level)	99.0566 %	
Mean rel. region size (0.95 level)	99.5283 %	
Total Number of Instances	106	

=== Detailed Accuracy By Class ===

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Distrust	0,640	0,250	0,696	0,640	0,667	0,393	0,733	0,696
Trust	0,750	0,360	0,700	0,750	0,724	0,393	0,733	0,739
Weighted Avg.	0,698	0,308	0,698	0,698	0,697	0,393	0,733	0,719

==== Confusion Matrix ====

```
a b <-- classified as
32 18 | a = Distrust
14 42 | b = Trust
```

Appendix 3 – Neural Network Classifier output in WEKA

==== Run information ====

```
Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -
E 20 -H a
Relation: FPSnr3-weka.filters.unsupervised.attribute.Remove-R14,16-17
Instances: 106
Attributes: 14
    Gender in nr.
    Age
    Years of use
    nr of people sharing
    Favourite SN in nr.
    Imp.News Sr. Nr
    Forwarding untrusted sr. In nr
    S vs T media in nr
    Blocking a pr. In nr
    Trust in previous posts In nr
    Use > 1 SN in nr
    nr of followers in nr.
    Field of study in nr.
    Trust in nominal
Test mode: evaluate on training data
```

==== Classifier model (full training set) ====

Sigmoid Node 0

```
Inputs  Weights
Threshold -2.5485507270834153
Node 2 -5.988037389418108
Node 3 6.42830986497764
Node 4 2.800728884979555
Node 5 -6.909232900989926
Node 6 -3.880168799814336
Node 7 8.114222676306907
Node 8 8.949745082890903
```

Sigmoid Node 1

```
Inputs  Weights
Threshold 2.5487280473022964
Node 2 5.988738140410537
Node 3 -6.428862324745234
Node 4 -2.8013094652497683
Node 5 6.909247410147893
```

Node 6 3.880347351784651
Node 7 -8.114397961013212
Node 8 -8.950090070754994

Sigmoid Node 2

Inputs Weights
Threshold 2.086468988401353
Attrib Gender in nr. -2.276490664254665
Attrib Age -0.2955578635132129
Attrib Years of use 1.7150948512003954
Attrib nr of people sharing 2.704618879425661
Attrib Favourite SN in nr. -0.24949550701569714
Attrib Imp.News Sr. Nr -1.4388283233895651
Attrib Forwarding untrusted sr. In nr 1.5946983679767577
Attrib S vs T media in nr 1.172262382898669
Attrib Blocking a pr. In nr -1.2719784277440545
Attrib Trust in previous posts In nr -3.2365394217422185
Attrib Use > 1 SN in nr -3.5945833031198444
Attrib nr of followers in nr. 1.4927777149089476
Attrib Field of study in nr. -3.1768697516864575

Sigmoid Node 3

Inputs Weights
Threshold -3.0175107607033547
Attrib Gender in nr. -5.201718805179545
Attrib Age 1.5384778804081896
Attrib Years of use 1.2446595624547325
Attrib nr of people sharing -6.9948397140680765
Attrib Favourite SN in nr. 2.6335493919976796
Attrib Imp.News Sr. Nr 2.8037697250518665
Attrib Forwarding untrusted sr. In nr 0.8260953317773042
Attrib S vs T media in nr -0.3529347892561964
Attrib Blocking a pr. In nr 0.5457906470638045
Attrib Trust in previous posts In nr 0.13459301610612717
Attrib Use > 1 SN in nr 4.111222604316842
Attrib nr of followers in nr. 0.926240253256749
Attrib Field of study in nr. 2.396159292161154

Sigmoid Node 4

Inputs Weights
Threshold -1.0782642318931728
Attrib Gender in nr. 2.1676037161808654
Attrib Age 0.3741376001486107
Attrib Years of use -1.4357084099131345
Attrib nr of people sharing 0.9192947616145641
Attrib Favourite SN in nr. 0.1276204685832617
Attrib Imp.News Sr. Nr -0.7999976589852235
Attrib Forwarding untrusted sr. In nr 1.1300940390835736
Attrib S vs T media in nr -1.2465833679880838
Attrib Blocking a pr. In nr -1.3201557293508486
Attrib Trust in previous posts In nr 0.21089567129063397
Attrib Use > 1 SN in nr -3.2483072445413304
Attrib nr of followers in nr. 0.36816534572752524

Attrib Field of study in nr. 1.3224989836592758
Sigmoid Node 5
Inputs Weights
Threshold 0.7676027050293622
Attrib Gender in nr. -2.7579591361350992
Attrib Age -3.1258958864703543
Attrib Years of use 4.784087065678835
Attrib nr of people sharing 7.230586124271508
Attrib Favourite SN in nr. -2.3405411285545155
Attrib Imp.News Sr. Nr -4.796142910892761
Attrib Forwarding untrusted sr. In nr 0.788981607159444
Attrib S vs T media in nr 1.8877240660175214
Attrib Blocking a pr. In nr 1.049778921233178
Attrib Trust in previous posts In nr -6.892724980413451
Attrib Use > 1 SN in nr -3.662518230044334
Attrib nr of followers in nr. 2.8613705351738226
Attrib Field of study in nr. -4.3021427474322085

Sigmoid Node 6
Inputs Weights
Threshold 1.3463653626045402
Attrib Gender in nr. 0.2944933620378113
Attrib Age -3.317331663083773
Attrib Years of use -2.4840353474246646
Attrib nr of people sharing -3.350642469978447
Attrib Favourite SN in nr. -1.87372048678515
Attrib Imp.News Sr. Nr -3.2517394294634445
Attrib Forwarding untrusted sr. In nr -1.1192807160634304
Attrib S vs T media in nr 0.7928807995824618
Attrib Blocking a pr. In nr -0.15046006104303872
Attrib Trust in previous posts In nr 0.6792115213167451
Attrib Use > 1 SN in nr -1.7531995218077014
Attrib nr of followers in nr. -1.67179914992699
Attrib Field of study in nr. 3.7838227812549325

Sigmoid Node 7
Inputs Weights
Threshold -6.771714820807276
Attrib Gender in nr. -0.00188081530265916
Attrib Age 2.7965134330063433
Attrib Years of use 2.5791326461987403
Attrib nr of people sharing 4.60059983688403
Attrib Favourite SN in nr. 0.766694300933294
Attrib Imp.News Sr. Nr -1.0920265646889011
Attrib Forwarding untrusted sr. In nr 0.39514389705638964
Attrib S vs T media in nr -5.033851268451308
Attrib Blocking a pr. In nr 0.47279304430725994
Attrib Trust in previous posts In nr -5.122796484072341
Attrib Use > 1 SN in nr 2.8503045718826163
Attrib nr of followers in nr. 6.715750960729068
Attrib Field of study in nr. 2.5819266518838973

Sigmoid Node 8

Inputs Weights

Threshold 1.2766027117653436
 Attrib Gender in nr. -4.094830871638469
 Attrib Age -4.104446377759735
 Attrib Years of use 0.29544182308743344
 Attrib nr of people sharing 1.113804782648516
 Attrib Favourite SN in nr. 3.184951713270853
 Attrib Imp.News Sr. Nr -4.121782080313684
 Attrib Forwarding untrusted sr. In nr 2.3838623605142586
 Attrib S vs T media in nr -0.4053337752168868
 Attrib Blocking a pr. In nr -0.19476620867873595
 Attrib Trust in previous posts In nr 4.7020621717427105
 Attrib Use > 1 SN in nr -5.141237383736017
 Attrib nr of followers in nr. -2.6004141089228785
 Attrib Field of study in nr. -5.75356083260466

Class Distrust

Input
 Node 0

Class Trust

Input
 Node 1

Time taken to build model: 0.56 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	102	96.2264 %
Incorrectly Classified Instances	4	3.7736 %
Kappa statistic	0.9241	
Mean absolute error	0.0627	
Root mean squared error	0.184	
Relative absolute error	12.5704 %	
Root relative squared error	36.8581 %	
Coverage of cases (0.95 level)	97.1698 %	
Mean rel. region size (0.95 level)	58.4906 %	
Total Number of Instances	106	

=== Detailed Accuracy By Class ===

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Distrust	0,940	0,018	0,979	0,940	0,959	0,925	0,945	0,931
Trust	0,982	0,060	0,948	0,982	0,965	0,925	0,945	0,886
Weighted Avg.	0,962	0,040	0,963	0,962	0,962	0,925	0,945	0,907

==== Confusion Matrix ====

a b <-- classified as
47 3 | a = Distrust
1 55 | b = Trust

Appendix 4 Rules PART output in WEKA

==== Run information ====

Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation: FPSnr3-weka.filters.unsupervised.attribute.Remove-R14,16-17
Instances: 106
Attributes: 14
Gender in nr.
Age
Years of use
nr of people sharing
Favourite SN in nr.
Imp.News Sr. Nr
Forwarding untrusted sr. In nr
S vs T media in nr
Blocking a pr. In nr
Trust in previous posts In nr
Use > 1 SN in nr
nr of followers in nr.
Field of study in nr.
Trust in nominal
Test mode: evaluate on training data

==== Classifier model (full training set) ====

PART decision list

Trust in previous posts In nr <= 0 AND
Imp.News Sr. Nr <= 2 AND
Use > 1 SN in nr > 0: Trust (14.0/1.0)

Gender in nr. <= 0 AND
Use > 1 SN in nr > 0 AND
nr of followers in nr. > 0: Distrust (10.0/1.0)

Favourite SN in nr. > 0 AND
Years of use <= 7 AND
Blocking a pr. In nr > 0: Distrust (8.0)

Favourite SN in nr. > 0 AND
Trust in previous posts In nr > 0: Trust (5.0)

Forwarding untrusted sr. In nr > 0: Distrust (6.0/2.0)

Field of study in nr. <= 0 AND
Years of use <= 2: Distrust (3.0)

Field of study in nr. <= 0 AND
Blocking a pr. In nr <= 0 AND
Use > 1 SN in nr <= 0 AND
nr of people sharing <= 4 AND
Age <= 26: Distrust (3.0/1.0)

Field of study in nr. <= 0 AND
Blocking a pr. In nr > 0 AND
Age <= 30 AND
Trust in previous posts In nr > 0 AND
S vs T media in nr <= 0 AND
nr of followers in nr. > 0: Trust (9.0/1.0)

Field of study in nr. <= 0 AND
Blocking a pr. In nr <= 0: Trust (6.0)

Blocking a pr. In nr <= 0 AND
Imp.News Sr. Nr <= 0: Distrust (4.0)

Blocking a pr. In nr > 0 AND
nr of followers in nr. <= 0 AND
Gender in nr. > 0 AND
Field of study in nr. <= 0 AND
Imp.News Sr. Nr <= 1 AND
Age > 25: Distrust (4.0/1.0)

Blocking a pr. In nr > 0 AND
nr of followers in nr. <= 0 AND
Gender in nr. > 0: Trust (7.0)

Blocking a pr. In nr > 0 AND
Use > 1 SN in nr > 0 AND
Field of study in nr. <= 0 AND
nr of people sharing > 1: Distrust (4.0)

Blocking a pr. In nr > 0 AND
Field of study in nr. <= 0 AND
Gender in nr. > 0: Distrust (5.0/2.0)

Field of study in nr. > 0 AND
Blocking a pr. In nr > 0 AND
Use > 1 SN in nr > 0 AND
nr of people sharing <= 3: Distrust (5.0/1.0)

nr of followers in nr. > 0 AND

S vs T media in nr <= 0 AND
Years of use <= 8: Distrust (5.0/1.0)

: Trust (8.0)

Number of Rules : 17

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	95	89.6226 %
Incorrectly Classified Instances	11	10.3774 %
Kappa statistic	0.7933	
Mean absolute error	0.156	
Root mean squared error	0.2793	
Relative absolute error	31.2964 %	
Root relative squared error	55.9448 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	78.7736 %	
Total Number of Instances	106	

=== Detailed Accuracy By Class ===

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
	0,960	0,161	0,842	0,960	0,897	0,800	0,956	0,938
	0,839	0,040	0,959	0,839	0,895	0,800	0,956	0,955
Weighted Avg.	0,896	0,097	0,904	0,896	0,896	0,800	0,956	0,947

=== Confusion Matrix ===

```
a b <-- classified as
48 2 | a = Distrust
9 47 | b = Trust
```

Appendix 5 K-Means Clustering seed value =10 and Distance Function = Euclidean Distance

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: FPSnr3-weka.filters.unsupervised.attribute.Remove-R15-17

Instances: 106

Attributes: 14

- Gender in nr.
- Age
- Years of use
- nr of people sharing
- Favourite SN in nr.
- Imp.News Sr. Nr
- Forwarding untrusted sr. In nr
- S vs T media in nr
- Blocking a pr. In nr
- Trust in previous posts In nr
- Use > 1 SN in nr
- nr of followers in nr.
- Field of study in nr.
- Trust in SN binary

Test mode: split 75% train, remainder test

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 7

Within cluster sum of squared errors: 210.7998591873908

Initial starting points (random):

Cluster 0: 0,25,6,1,0,3,0,0,1,1,1,0,0,1

Cluster 1: 1,29,5,5,0,0,0,0,1,1,1,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#		
	Full Data (106)	0 (45)	1 (61)
Gender in nr.	0.6415	0.4667	0.7705
Age	27.2075	27.0444	27.3279
Years of use	5.6792	5.6444	5.7049
nr of people sharing	2.8302	2.6444	2.9672
Favourite SN in nr.	0.2642	0.3333	0.2131
Imp.News Sr. Nr	0.9623	1.0222	0.918
Forwarding untrusted sr. In nr	0.066	0	0.1148
S vs T media in nr	0.1792	0.2667	0.1148
Blocking a pr. In nr	0.6981	0.4667	0.8689
Trust in previous posts In nr	0.7642	0.4667	0.9836
Use > 1 SN in nr	0.717	0.5111	0.8689

nr of followers in nr.	0.6226	0.8	0.4918
Field of study in nr.	0.3774	0.3556	0.3934
Trust in SN binary	0.5283	0.6889	0.4098

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on test split ===

kMeans

=====

Number of iterations: 11

Within cluster sum of squared errors: 153.5458461247747

Initial starting points (random):

Cluster 0: 1,28,3,2,0,2,0,0,1,1,1,1,0,1

Cluster 1: 1,24,5,1,0,0,0,0,0,1,0,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#		
	Full Data (79)	0 (52)	1 (27)
Gender in nr.	0.6329	0.7692	0.3704
Age	27.5063	27.8269	26.8889
Years of use	5.6203	5.6538	5.5556
nr of people sharing	2.7848	3	2.3704
Favourite SN in nr.	0.2785	0.1731	0.4815
Imp.News Sr. Nr	0.8861	0.8846	0.8889
Forwarding untrusted sr. In nr	0.038	0.0192	0.0741
S vs T media in nr	0.1899	0.2115	0.1481
Blocking a pr. In nr	0.6962	0.9615	0.1852
Trust in previous posts In nr	0.7342	0.75	0.7037
Use > 1 SN in nr	0.6962	0.7885	0.5185
nr of followers in nr.	0.6203	0.6731	0.5185
Field of study in nr.	0.3924	0.2885	0.5926
Trust in SN binary	0.4937	0.5962	0.2963

Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0 20 (74%)
1 7 (26%)

Appendix 6 K-Means Clustering seed value =50 and Distance Function = Euclidean Distance

==== Run information ====

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 50

Relation: FPSnr3-weka.filters.unsupervised.attribute.Remove-R15-17

Instances: 106

Attributes: 14

Gender in nr.
Age
Years of use
nr of people sharing
Favourite SN in nr.
Imp.News Sr. Nr
Forwarding untrusted sr. In nr
S vs T media in nr
Blocking a pr. In nr
Trust in previous posts In nr
Use > 1 SN in nr
nr of followers in nr.
Field of study in nr.
Trust in SN binary

Test mode: evaluate on training data

==== Clustering model (full training set) ====

kMeans

=====

Number of iterations: 5

Within cluster sum of squared errors: 207.58300796332045

Initial starting points (random):

Cluster 0: 0,29,7,5,0,1,0,0,1,1,1,0,1,1

Cluster 1: 0,28,11,5,2,1,0,0,0,1,1,0,0,1

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#		
	Full Data (106)	0 (74)	1 (32)
Gender in nr.	0.6415	0.6757	0.5625
Age	27.2075	27.3378	26.9063
Years of use	5.6792	5.7297	5.5625
nr of people sharing	2.8302	2.973	2.5
Favourite SN in nr.	0.2642	0.2162	0.375
Imp.News Sr. Nr	0.9623	1	0.875
Forwarding untrusted sr. In nr	0.066	0.0541	0.0938
S vs T media in nr	0.1792	0.1892	0.1563
Blocking a pr. In nr	0.6981	1	0
Trust in previous posts In nr	0.7642	0.8108	0.6563
Use > 1 SN in nr	0.717	0.7973	0.5313
nr of followers in nr.	0.6226	0.6081	0.6563
Field of study in nr.	0.3774	0.3378	0.4688
Trust in SN binary	0.5283	0.5405	0.5

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 74 (70%)
1 32 (30%)

Appendix 7 K-Means Clustering seed value =100 and Distance Function = Euclidean Distance

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 100

Relation: FPSnr3

Instances: 106

Attributes: 17

Gender in nr.
Age
Years of use
nr of people sharing
Favourite SN in nr.
Imp.News Sr. Nr
Forwarding untrusted sr. In nr

S vs T media in nr
 Blocking a pr. In nr
 Trust in previous posts In nr
 Use > 1 SN in nr
 nr of followers in nr.
 Field of study in nr.
 Trust in SN binary
 Trust in nominal
 Trust in SN
 Field of study

Test mode: evaluate on training data

==== Clustering model (full training set) ====

kMeans

=====

Number of iterations: 3

Within cluster sum of squared errors: 292.4611641399416

Initial starting points (random):

Cluster 0: 0,27,5,1,0,1,0,1,0,1,0,0,0,1,Trust,3,Mathematics

Cluster 1: 0,24,2,5,0,2,0,0,1,1,1,1,1,0,Distrust,2,'Comparative education'

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#		
	Full Data (106)	0 (56)	1 (50)
Gender in nr.	0.6415	0.6964	0.58
Age	27.2075	27.0357	27.4
Years of use	5.6792	5.8214	5.52
nr of people sharing	2.8302	2.8571	2.8
Favourite SN in nr.	0.2642	0.1964	0.34
Imp.News Sr. Nr	0.9623	0.8929	1.04
Forwarding untrusted sr. In nr	0.066	0.0536	0.08
S vs T media in nr	0.1792	0.2321	0.12
Blocking a pr. In nr	0.6981	0.7143	0.68
Trust in previous posts In nr	0.7642	0.6786	0.86
Use > 1 SN in nr	0.717	0.6964	0.74
nr of followers in nr.	0.6226	0.6607	0.58
Field of study in nr.	0.3774	0.3214	0.44
Trust in SN binary	0.5283	1	0
Trust in nominal	Trust	Trust	Distrust

Trust in SN	2.4906	3.3214	1.56
Field of study	Computer Science	Computer Science	Computer Science

Time taken to build model (full training data) : 0.09 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	56 (53%)
1	50 (47%)

Appendix 8 K-Means Clustering seed value =1000 and Distance Function = Manhattan Distance

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.ManhattanDistance -R first-last" -I 500 -num-slots 1 -S 1000

Relation: FPSnr3

Instances: 106

Attributes: 17

- Gender in nr.
- Age
- Years of use
- nr of people sharing
- Favourite SN in nr.
- Imp.News Sr. Nr
- Forwarding untrusted sr. In nr
- S vs T media in nr
- Blocking a pr. In nr
- Trust in previous posts In nr
- Use > 1 SN in nr
- nr of followers in nr.
- Field of study in nr.
- Trust in SN binary
- Trust in nominal
- Trust in SN
- Field of study

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 2
 Sum of within cluster distances: 438.52857142857147

Initial starting points (random):

Cluster 0: 0,27,9,5,0,3,0,0,1,1,1,1,1,0,Distrust,1,'public administration'
 Cluster 1: 0,29,7,5,0,1,0,0,1,1,1,0,1,1,Trust,3,Journalism

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#		
	Full Data (106)	0 (50)	1 (56)
Gender in nr.	1	1	1
Age	28	28	28
Years of use	5	5	5
nr of people sharing	2	2	2
Favourite SN in nr.	0	0	0
Imp.News Sr. Nr	0.5	0.5	0.5
Forwarding untrusted sr. In nr	0	0	0
S vs T media in nr	0	0	0
Blocking a pr. In nr	1	1	1
Trust in previous posts In nr	1	1	1
Use > 1 SN in nr	1	1	1
nr of followers in nr.	1	1	1
Field of study in nr.	0	0	0
Trust in SN binary	1	0	1
Trust in nominal	Trust	Distrust	Trust
Trust in SN	3	2	3
Field of study	Computer Science	Computer Science	Computer Science

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 50 (47%)
 1 56 (53%)

Appendix 9 Linear regression output in Minitab

Regression Analysis: Trust in SN versus Gender in nr.; Age; ...

The regression equation is

Trust in SN = 4,03 + 0,200 Gender in nr. - 0,0695 Age + 0,109 Years of use
+ 0,0034 nr of people sharing - 0,214 Favourite SN in nr.
- 0,0622 Imp.News Sr. Nr + 0,288 Forwarding untrusted sr.
In nr
+ 0,122 S vs T media in nr + 0,065 Blocking a pr. In nr
- 0,181 Trust in previous posts In nr - 0,306 Use > 1 SN
in nr
+ 0,233 nr of followers in nr. - 0,428 Field of study in nr.

Predictor	Coef	SE Coef	T	P
Constant	4,030	1,025	3,93	0,000
Gender in nr.	0,1995	0,2415	0,83	0,411
Age	-0,06946	0,03477	-2,00	0,049
Years of use	0,10875	0,05119	2,12	0,036
nr of people sharing	0,00339	0,06727	0,05	0,960
Favourite SN in nr.	-0,2139	0,1853	-1,15	0,251
Imp.News Sr. Nr	-0,06224	0,09674	-0,64	0,522
Forwarding untrusted sr. In nr	0,2876	0,4797	0,60	0,550
S vs T media in nr	0,1217	0,3106	0,39	0,696
Blocking a pr. In nr	0,0648	0,2586	0,25	0,803
Trust in previous posts In nr	-0,1815	0,2779	-0,65	0,515
Use > 1 SN in nr	-0,3062	0,2619	-1,17	0,245
nr of followers in nr.	0,2330	0,2277	1,02	0,309
Field of study in nr.	-0,4275	0,2344	-1,82	0,071

S = 1,08168 R-Sq = 17,5% R-Sq(adj) = 5,9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	13	22,848	1,758	1,50	0,132
Residual Error	92	107,643	1,170		
Total	105	130,491			

Source	DF	Seq SS
Gender in nr.	1	1,306
Age	1	3,516
Years of use	1	5,899
nr of people sharing	1	0,527
Favourite SN in nr.	1	1,567
Imp.News Sr. Nr	1	0,295
Forwarding untrusted sr. In nr	1	0,328
S vs T media in nr	1	1,255
Blocking a pr. In nr	1	0,003
Trust in previous posts In nr	1	1,019
Use > 1 SN in nr	1	2,290
nr of followers in nr.	1	0,952

Field of study in nr. 1 3,892

Unusual Observations

Obs	Gender in nr.	Trust in SN	Fit	SE Fit	Residual	St Resid
3	1,00	5,000	2,811	0,494	2,189	2,27R
5	1,00	0,000	2,172	0,462	-2,172	-2,22R
33	1,00	0,000	2,173	0,286	-2,173	-2,08R
79	0,00	5,000	2,321	0,497	2,679	2,79R
87	0,00	0,000	2,505	0,417	-2,505	-2,51R
91	1,00	0,000	2,158	0,397	-2,158	-2,14R

R denotes an observation with a large standardized residual.

Normplot of Residuals for Trust in SN

Residuals vs Fits for Trust in SN

Residual Histogram for Trust in SN

Probability Plot of RESI1

Retrieving worksheet from file: 'C:\Users\Sim - One\Documents\FPSnr.xlsx'
Worksheet was saved on 24.05.2014

Appendix 10 Binary Logistic Regression in Minitab

Binary Logistic Regression: Trust in SN versus Gender in nr; Age; ...

Link Function: Logit

Response Information

Variable	Value	Count
Trust in SN binary	1	56 (Event)
	0	50
	Total	106

Logistic Regression Table

Odds Predictor Ratio	Coef	SE Coef	Z	P
Constant	2,17191	2,06805	1,05	0,294
Gender in nr. 1,69	0,527662	0,476122	1,11	0,268

Age	-0,0859213	0,0703115	-1,22	0,222
0,92				
Years of use	0,102597	0,102593	1,00	0,317
1,11				
nr of people sharing	0,0938015	0,134129	0,70	0,484
1,10				
Favourite SN in nr.	-0,481409	0,389553	-1,24	0,217
0,62				
Imp.News Sr. Nr	-0,146963	0,190598	-0,77	0,441
0,86				
Forwarding untrusted sr. In nr	-1,04683	0,936633	-1,12	0,264
0,35				
S vs T media in nr	0,742731	0,644067	1,15	0,249
2,10				
Blocking a pr. In nr	0,0338852	0,520169	0,07	0,948
1,03				
Trust in previous posts In nr	-0,846473	0,575227	-1,47	0,141
0,43				
Use > 1 SN in nr	0,0245806	0,538290	0,05	0,964
1,02				
nr of followers in nr.	0,329272	0,446629	0,74	0,461
1,39				
Field of study in nr.	-0,724665	0,476035	-1,52	0,128
0,48				

Predictor	95% CI	
	Lower	Upper
Constant		
Gender in nr.	0,67	4,31
Age	0,80	1,05
Years of use	0,91	1,35
nr of people sharing	0,84	1,43
Favourite SN in nr.	0,29	1,33
Imp.News Sr. Nr	0,59	1,25
Forwarding untrusted sr. In nr	0,06	2,20
S vs T media in nr	0,59	7,43
Blocking a pr. In nr	0,37	2,87
Trust in previous posts In nr	0,14	1,32
Use > 1 SN in nr	0,36	2,94
nr of followers in nr.	0,58	3,34
Field of study in nr.	0,19	1,23

Log-Likelihood = -66,092

Test that all slopes are zero: G = 14,423, DF = 13, P-Value = 0,345

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	107,315	92	0,131
Deviance	132,184	92	0,004
Hosmer-Lemeshow	5,691	8	0,682

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Value	Group										Total
	1	2	3	4	5	6	7	8	9	10	
1											
Obs	1	6	4	6	4	4	6	7	8	10	56
Exp	2,4	3,8	4,0	4,8	5,5	5,4	6,5	6,3	8,1	9,3	
0											
Obs	9	5	6	5	7	6	5	3	3	1	50
Exp	7,6	7,2	6,0	6,2	5,5	4,6	4,5	3,7	2,9	1,7	
Total	10	11	10	11	11	10	11	10	11	11	106

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	1960	70,0	Somers' D	0,40
Discordant	829	29,6	Goodman-Kruskal Gamma	0,41
Ties	11	0,4	Kendall's Tau-a	0,20
Total	2800	100,0		

Delta Chi-Square versus P

Delta Chi-Square versus Hi

Appendix 11 Logistic regression output in R

```
Call: glm(formula = Trust.in.SN.binary ~ Age + Years.of.use +
Gender.in.nr. +
nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
Blocking.a..pr..In.nr +
Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
nr.of.followers.in.nr. +
Field.of.study.in.nr., family = binomial, data = SN)
```

Coefficients:

(Intercept)	Age
2.17191	-0.08592
Years.of.use	Gender.in.nr.
0.10260	0.52766
nr.of.people.sharing	Favourite.SN.in.nr.
0.09380	-0.48141
Imp.News.Sr..Nr	Forwarding.untrusted.sr..In.nr
-0.14696	-1.04683
S.vs.T.media.in.nr	Blocking.a..pr..In.nr
0.74273	0.03389
Trust.in.previous.posts.In.nr	Use...1.SN.in.nr
-0.84647	0.02458
nr.of.followers.in.nr.	Field.of.study.in.nr.
0.32927	-0.72466

Degrees of Freedom: 105 Total (i.e. Null); 92 Residual

Null Deviance: 146.6

Residual Deviance: 132.2 AIC: 160.2

> summary (model)

```
Call:
glm(formula = Trust.in.SN.binary ~ Age + Years.of.use + Gender.in.nr. +
     nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
     Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
Blocking.a..pr..In.nr +
     Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
nr.of.followers.in.nr. +
     Field.of.study.in.nr., family = binomial, data = SN)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2017	-1.0734	0.5227	1.0624	1.6166

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.17191	2.06798	1.050	0.294
Age	-0.08592	0.07031	-1.222	0.222
Years.of.use	0.10260	0.10259	1.000	0.317
Gender.in.nr.	0.52766	0.47611	1.108	0.268
nr.of.people.sharing	0.09380	0.13412	0.699	0.484
Favourite.SN.in.nr.	-0.48141	0.38954	-1.236	0.217
Imp.News.Sr..Nr	-0.14696	0.19059	-0.771	0.441
Forwarding.untrusted.sr..In.nr	-1.04683	0.93660	-1.118	0.264
S.vs.T.media.in.nr	0.74273	0.64404	1.153	0.249
Blocking.a..pr..In.nr	0.03389	0.52015	0.065	0.948
Trust.in.previous.posts.In.nr	-0.84647	0.57520	-1.472	0.141
Use...1.SN.in.nr	0.02458	0.53827	0.046	0.964
nr.of.followers.in.nr.	0.32927	0.44662	0.737	0.461
Field.of.study.in.nr.	-0.72466	0.47602	-1.522	0.128

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 146.61 on 105 degrees of freedom
Residual deviance: 132.18 on 92 degrees of freedom
AIC: 160.18

Number of Fisher Scoring iterations: 3

>

Appendix 12 Poisson Regression Output in R

> model

```
Call: glm(formula = Trust.in.SN.binary ~ Age + Years.of.use +
     Gender.in.nr. +
     nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
     Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
Blocking.a..pr..In.nr +
     Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
nr.of.followers.in.nr. +
     Field.of.study.in.nr., family = poisson, data = SN)
```

Coefficients:

(Intercept)	Age
0.07583	-0.03529
Years.of.use	Gender.in.nr.

	0.04185		0.24706
nr.of.people.sharing		Favourite.SN.in.nr.	
	0.04104		-0.22200
Imp.News.Sr..Nr		Forwarding.untrusted.sr..In.nr	
	-0.04950		-0.43945
S.vs.T.media.in.nr		Blocking.a..pr..In.nr	
	0.25311		-0.01243
Trust.in.previous.posts.In.nr		Use...1.SN.in.nr	
	-0.28005		0.03091
nr.of.followers.in.nr.		Field.of.study.in.nr.	
	0.13560		-0.30331

Degrees of Freedom: 105 Total (i.e. Null); 92 Residual
Null Deviance: 71.47
Residual Deviance: 65.33 AIC: 205.3
> summary (model)

Call:
glm(formula = Trust.in.SN.binary ~ Age + Years.of.use + Gender.in.nr. +
nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
Blocking.a..pr..In.nr +
Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
nr.of.followers.in.nr. +
Field.of.study.in.nr., family = poisson, data = SN)

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.48179	-0.94791	0.03353	0.54419	0.99993

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.07583	1.31637	0.058	0.954
Age	-0.03529	0.04413	-0.800	0.424
Years.of.use	0.04185	0.06517	0.642	0.521
Gender.in.nr.	0.24706	0.31831	0.776	0.438
nr.of.people.sharing	0.04104	0.08572	0.479	0.632
Favourite.SN.in.nr.	-0.22200	0.26478	-0.838	0.402
Imp.News.Sr..Nr	-0.04950	0.12665	-0.391	0.696
Forwarding.untrusted.sr..In.nr	-0.43945	0.66341	-0.662	0.508
S.vs.T.media.in.nr	0.25311	0.37062	0.683	0.495
Blocking.a..pr..In.nr	-0.01243	0.33799	-0.037	0.971
Trust.in.previous.posts.In.nr	-0.28005	0.33199	-0.844	0.399
Use...1.SN.in.nr	0.03091	0.33104	0.093	0.926
nr.of.followers.in.nr.	0.13560	0.29493	0.460	0.646
Field.of.study.in.nr.	-0.30331	0.31078	-0.976	0.329

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 71.466 on 105 degrees of freedom
Residual deviance: 65.328 on 92 degrees of freedom
AIC: 205.33

Number of Fisher Scoring iterations: 5

> plot (model)
Waiting to confirm page change...
Waiting to confirm page change...

Waiting to confirm page change...
Waiting to confirm page change...

Appendix 13 Negative Binomial Regression Output R

Call:
glm.nb(formula = Trust.in.SN.binary ~ Age + Years.of.use +
Gender.in.nr. +
nr.of.people.sharing + Favourite.SN.in.nr. + Imp.News.Sr..Nr +
Forwarding.untrusted.sr..In.nr + S.vs.T.media.in.nr +
Blocking.a..pr..In.nr +
Trust.in.previous.posts.In.nr + Use...1.SN.in.nr +
nr.of.followers.in.nr. +
Field.of.study.in.nr., data = SN, init.theta = 19110.19472,
link = log)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.48178	-0.94791	0.03353	0.54417	0.99992

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.07582	1.31640	0.058	0.954
Age	-0.03529	0.04413	-0.800	0.424
Years.of.use	0.04185	0.06517	0.642	0.521
Gender.in.nr.	0.24706	0.31832	0.776	0.438
nr.of.people.sharing	0.04104	0.08573	0.479	0.632
Favourite.SN.in.nr.	-0.22200	0.26479	-0.838	0.402
Imp.News.Sr..Nr	-0.04950	0.12665	-0.391	0.696
Forwarding.untrusted.sr..In.nr	-0.43945	0.66342	-0.662	0.508
S.vs.T.media.in.nr	0.25311	0.37063	0.683	0.495
Blocking.a..pr..In.nr	-0.01242	0.33800	-0.037	0.971
Trust.in.previous.posts.In.nr	-0.28005	0.33199	-0.844	0.399
Use...1.SN.in.nr	0.03091	0.33105	0.093	0.926
nr.of.followers.in.nr.	0.13560	0.29494	0.460	0.646
Field.of.study.in.nr.	-0.30331	0.31078	-0.976	0.329

(Dispersion parameter for Negative Binomial(19110.19) family taken to be 1)

The regression equation is

Trust in SN = 0.07582 + 0.24706 Gender in nr. - 0.03529 Age
+ 0.0419 Years of use + 0.0411 nr of people sharing
- 0.222 Favorite SN in nr. - 0.0495 Imp.News Sr Nr
- 0.43945 forwarding un trusted sr. In nr
+ 0.253 SvsT media in nr + 0.0124 Blocking pr.In nr
- 0.28 Trust in previous posts In nr
+ 0.0309 Use>1 SN in nr + 0.136 nr of followers nr.
- 0.3033 Field of study in nr.

Null deviance: 71.464 on 105 degrees of freedom

Residual deviance: 65.327 on 92 degrees of freedom
AIC: 207.33

Number of Fisher Scoring iterations: 1

Theta: 19110
Std. Err.: 252042
Warning while fitting theta: iteration limit reached
2 x log-likelihood: -177.33