Shahjalal University of Science and Technology
Department of Computing Science and Engineering

Sentiment Analysis of Bengali Text Using Deep Learning

Abdullah Aziz Sharfuddin
Reg. No.: 2014331011
4^{th} year, 1^{st} Semester

Md. Nafis Tihami
Reg. No.: 2014331026
4^{th} year, 1^{st} Semester

Department of Computer Science and Engineering

Supervisor
Md. Saiful Islam
Assistant Professor
Department of Computer Science and Engineering

14th July, 2018

Sentiment Analysis of Bengali Text Using Deep Learning

A Thesis Submitted to the

Department of Computing Science and Engineering

Shahjalal University of Science and Technology

Sylhet - 3114, Bangladesh

in partial fulfillment of the requirements for the degree of

B.Sc.(Engg.) in Computer Science and Engineering

By

Abdullah Aziz Sharfuddin

Reg. No.: 2014331011

4^{th} year, 1^{st} Semester

Md. Nafis Tihami

Reg. No.: 2014331026

4^{th} year, 1^{st} Semester

Department of Computer Science and Engineering

Supervisor

Md. Saiful Islam

Assistant Professor

Department of Computer Science and Engineering

14th July, 2018

Recommendation Letter from Thesis Supervisor

The thesis
entitled "Sentiment Analysis of Bengali Text Using Deep Learning"
submitted by the students

1. Abdullah Aziz Sharfuddin
2. Md. Nafis Tihami

is a record of research work carried out under my supervision. I, hereby, agree that the thesis can be submitted for examination.

Signature of the Supervisor:
Name of the Supervisor: Md. Saiful Islam
Date: 14th July, 2018

Certificate of Acceptance of the Thesis

The thesis

entitled "Sentiment Analysis of Bengali Text Using Deep Learning"

submitted by the students

1. Abdullah Aziz Sharfuddin
2. Md. Nafis Tihami

on 14th July, 2018

as part of the requirements of the course CSE-450, is being approved by the Department of Computer Science and Engineering as a partial fulfillment of the B.Sc.(Engg.) degree of the above students.

Head of the Dept.	Chairman, Exam. Committee	Supervisor
Dr Mohammad Reza Selim	M. Jahirul Islam, PhD, PEng	Md. Saiful Islam
Professor	Professor	Professor
Department of Computer	Department of Computer	Department of Computer
Science and Engineering	Science and Engineering	Science and Engineering

Abstract

Now a days people are more likely to share their feelings or opinions on the Internet more specifically on social media platforms like facebook, twitter etc. than to a real person. This is the new trend of speaking out. Their opinions on social media contains valuable insights such as over all polarity of a public election or the reactions on public events and movements or even whether people are liking the newly launched product of a brand and what not. As the number of social media users are already huge and increasing day by day, it has become very hard for someone to analyze these opinions or sentiments by hand. But these huge stream of data needs to be processed or classified for understanding the public reaction, their demands and what not. This why sentiment analysis has become a topic of great interest these days. In our thesis we present a way of sentiment classification of Bengali text using deep learning. We used RNN(recursive neural network) and LSTM(long short term memory) for classifying sentiments.

Keywords: Deep learning, RNN, LSTM, Sentiment Analysis, Classification, Bengali Text.

Acknowledgements

We would like to thank the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh for supporting this research. We are also thankful to the authors of previous works for their excellent paper.

Dedication

We would like to dedicate our research to our parents and to our younger brothers.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

A Online Social Networks (OSN) is an online platform which provides a user to interact and build a social relation as well as share their views and contents with other users who have similar interest or have a personal connections with the user. And one of the biggest platform of OSN is **Facebook** nowadays. The main idea behind this platform is to bring people together which represent too broad context to cover by a definition. User Generated content (UGC) are the backbone of a social network. There are a number of social networks other than Facebook which are popular among the peoples i.e Twitter, Instagram, LinkedIn and many others. Each of them have standalone features however in general all of them shares some common features. Now a days social networks are one of the easiest way to collect or share information. These are one of the most rapid information propagation system. If a user is considered as a node then those who are connected to the user will form a small network, and each user present in this network can potentially form another network with users connected to them which eventually leads to a much bigger network. Information originated from any of these small network travels rapidly to others networks through the common user between the networks. This simple form of information sharing however comes with a great flaw. As generating and sharing content is very easy in social network often there remains a question about the validity of the content. The users are responsible for validating the information before sharing it. Many users do not want to go to that extent instead they rely on the user from whom the information was propagated [1]. The cybercriminals take this

chance to share false or invalid information using a fake account as it is not possible to identify the owner of a fake account. In the upcoming section we will discuss deeply about fake accounts and its impact.

1.2 Fake Accounts

As mentioned earlier fake accounts are those accounts with credentials that do not represent the real identity of the user. Facebook has different kinds of accounts, In the upcoming chapter, we will discuss about it further. In our work, we will only consider accounts that reflect to be a human. These can be divided into following categories

1. **Inactive users :** To identify the validity of an account we need sufficient data. User account which does not have public post greater than a threshold[1] value does not provide the amount of data we need. This type of users are considered inactive users and possess no threat as they do not have enough influence on the network or its users.

2. **Active users :** Users who have public post and interactions more than a threshold value falls into this category. We consider these accounts to be influential accounts thus the validity of these accounts needs to be checked.

1.3 Reasons to Identify Fake Accounts

There are numbers of ways a fake account can create havoc in social media as well as in real life [2]. Some of the common phenomena are listed below-

1. **Spamming :** Cyber criminals and hackers have been utilizing spam for decades. In an online social network, spam is particularly effective at making an individual click on a link that would interest almost any normal user, most of the time they are simple and innocent-looking messages that you have to win a daily prize or your friend have invited you to play an online game. Then in order to claim your prize or play the game, you have to provide your credit card number or you have to log in to another site using your Facebook credentials. This way the hackers can get access to your private information [3].

[1] minimum 10 public post

2. **Cyberbullying :** Using of fake account in cyberbullying is also very common. It mainly takes place among the teenagers on social media and can leads to major criminal charges if it goes too far. There have been many cases reported that the victim of cyberbullying has committed suicide or was killed by a peer. Cyber-harassment or cyberstalking is a similar crime but involves adults.

3. **Identity Theft :** Personal information are literally floating around in social media nowadays and thus for cyber criminals it has become fairly easy to use this information to create an identical fake account of the real user and obtain private information from other users close to the real user [4].

4. **Defamation :** Defamation is another crime mainly done using a fake account. Propagating a false or negative statement about an individual or an entity is called defamation. A defamatory statement can harm the reputation of an individual or an organization severely. It is a punishable crime but a fake account makes it hard to identify the real offender.

5. **Stalking :** Stalking is a fairly common term used in online social media mostly as a joke. It means to visit someone's profile regularly. Most of the social media offers its users to block the unwanted friend. The blocked person then uses a fake account to stalk the user. Although it may seem like nothing more than annoying behavior, but it is a legitimate cause for concern in many cases.

6. **Harassment :** Nowadays Harassment in social media is happening quite often and is increasing day by day. Harassing messages, inappropriate comments, and other persistent behaviors are done using a fake account to remain invisible from the law.

Chapter 2

Facebook

2.1 Facebook

Facebook is an American online social media and social networking service company based in Menlo Park, California. The Facebook website was launched on February 4, 2004, by Mark Zuckerberg, along with fellow Harvard College students and roommates, Eduardo Saverin, Andrew McCollum, Dustin Moskovitz, and Chris Hughes[1].

Facebook provides a large number of functionalities. In our study we will limit our discussion among the followings -

1. User profile/Personal timeline
2. News Feed
3. Like/React
4. Following

2.2 Public and Private info

Users informations of a Facebook user can be divided into two categories.

[1]https://en.wikipedia.org/wiki/Facebook

1. **Public :** Profile information that has privacy settings as public, public post on user's timeline, content posted by the user in a public group, comment posted on a public post by the user falls into this category.

2. **Private :** Profile information that has privacy settings as private or only me, user's friend list, post on user's timeline which only the user and user's friends can see, content posted by the user in a closed group, comment posted on a private post by the user belongs to this category.

Due to Facebooks strict information sharing policy obtaining private information in not possible. In our study we will use only public information.

2.3 User classification

In our work for simplicity we have divided Facebook accounts into two group.

1. **Facebook Pages :** Facebook provide it's users to create profiles for a business, organization, or non-human entity. Such entities referred a page rather than a user profile. Although these accounts belongs to a human user these accounts are not subject to our study.

2. **Facebook Users :** Accounts which appear to be a human are member of these group and the main concern of our work. Our goal will be to isolate fake accounts from this group. There are mainly two types of fake profiles

   (a) **Redundant Accounts :** These accounts are maintained by a real user along with the user's primary account.

   (b) **Misused Accounts :** User profiles created to be used for purposes that violate Facebook's terms of service, such as spamming.

Chapter 3

Background Study

Before we jump into discuss our own method we should take a look at some basic machine learning methods and existing work. To begin with, Facebook has implemented in own immune system [5] to solve the problems regarding a fake profile by using a classifier. But the major drawback is that the authors do not provide this services to the users at large. Also there are many works which suggest that it does not solve the problems to the extent we expect.

3.1 Machine Learning methods :

We will start by reviewing some popular machine learning models and algorithms as these are being used in many previous works and we will also use some of the methods according to our criteria.

3.1.1 Support Vector Machines (SVM) :

Support vector machines are a type of classifier that separates the input data points into two class by finding the best hyper-plane between the two class. The best hyper-plane means that the data points of both class remains at maximum distance from the plane. Data points of the two class that are closest to the separating plane are called support vector.

3.1.2 Decision Tree :

Decision Trees are widely used for classification and regression model. It is a non-parametric

supervised learning method. It generates a tree incrementally using the smallest subsets of the dataset. The goal of decision tree is to create a model that learns simple decision rules inferred from the data features to predict the value of a unknown variable. The final decision tree consist of two types of nodes, leaf nodes and decision nodes. A decision nodes can have two or more branches depending on the data. A leaf node correspond to a decision or classification.

3.1.3 Naive Bayes Algorithm :

The Naive Bayes Classifier works based on Bayesian theorem and is more appropriate for input with high dimensionality. It's works using simple principle of probability. However despite of being simple in nature in some cases it can outperform many sophisticated classification method. Naive Bayes needs to scans the whole data set once and after that at any given time more training data can be added and probability of a data point being in a class gets updated accordingly.

3.1.4 Artificial Neural Network (ANN) :

Artificial neural network is a computational model inspired by structure and function of a biological neural network. It is generally composed of three layer, input layer, hidden layer and output layer. Hidden layer may contain one or more layer base on the model. There are also different structure of ANN based on different model. The biological neural network (central nervous system of animals to be specific) learns from experience. The artificial neural networks follows the same principal. The neurons in the network are interconnected to each other with a numeric value as weigh, it takes a large amount of data as input, computes the result using the weight and deliver the output. Artificial neural network is capable of adjusting the weight of the neuron itself based on the experience and that's why this model is much adaptive to input and capable of learning [6].

3.1.5 Deep Neural Network (DNN) :

Deep Neural Network is a type of Neural Network in which everything is alike as the earlier network like perceptrons network with a little bit different. In perceptrons network, there is an input, an output and at most one hidden layer. But using more than three layers is considered as 'deep learning'. In the deep neural network, we used more than one hidden layer, so that the internal hidden layers trains on a distinct feature list that is found from the previous layer.

As a result of that, the network can recognize a more complex pattern and the ultimate output is better than other neural network used earlier. Deep Neural Network uses Feature Hierarchy that helps the network to maintain abstraction among the layers and its corresponding activation. A deep neural network can manage to process more data than perceptron neural network.

3.1.6 K-Nearest Neighbors (KNN) :

K-nearest neighbors is a simple algorithm that classifies data points based on a similarity measure. Three kinds of distance function is useed to calculate similarity measure.

1. euclidean,

$$D = \sqrt{\sum_{n=1}^{k} (x_i - y_i)^2}$$

2. manhattan,

$$D = \sum_{n=1}^{k} |x_i - y_i|$$

3. minkawaski,

$$D = \left(\sum_{n=1}^{k} (|x_i - y_i|)^q \right)^{1/q}$$

These three distance measures are applicable only continuous variables. Hamming distance must be used when dealing with categorical variables. If there is a mixture of numerical and categorical variables in the dataset standardization of the numerical variables between 0 and 1 is needed.

3.1.7 Random Forest :

Random Forest is one of the most used supervised classification algorithm. It is flexible, easy to use and can be used for both classification and regression tasks.Most of the time it produces great result even without hyper-parameter tuning. The main idea of the algorithm is to create a forest with number of trees.

In general, the more the number of trees present in a forest the more robust it looks. Following the same principle in random forest algorithm the more trees there are the more accurate the result gets. In decision tree algorithm we use gini index or information gain to calculate root node.

But here instead of using information gain or gini index the process of finding the root node and splitting the feature nodes will happen randomly.

3.2 Related works :

This section contains short review about previous works to predict fake accounts in OSN. Although many of them has used **Twitter** and **LinkedIn** for their research [7–9], their applied methods can be used to other OSN's with a bit of modification.

3.2.1 Relative Comparison Approach:

In [10], facebook was also chosen as target OSN. They accumulated data of 777 users. Among them 229 accounts were previously marked as spam accounts (Fake accounts). The rest of the accounts were collected from the authors accounts and minimum 523 accounts were assumed as real accounts.

For feature selection, they manually selected 17 features that will be most useful while distinguishing real and fake users as far as possible. They tagged each feature with a feature name, feature description, justification of choosing the feature and measuring methodology. For example,
Name : `Average_PostLikes_Received`
Description : Average number of likes received by the user in his own post.
Justification : fake account is more likely to post and share false, malicious and spam messages which are expected to get low like counts.
Measuring Methodology : From the post of a users news feed the likes on that post can be collected.

Finally, they used 12 different supervised leaning classification algorithm and trained them separately. They evaluated the performance of each methods, and evaluated the accuracy of each method.

3.2.2 Statistical Approach:

In [11], temporal evaluation of a real users OSN's profile has been used to detect fake accounts. Here they first identified three features that has relation to a users social interaction and social

network graph properties. The main idea behind their approach is to find fake accounts based on these properties.

1. **Evolution of number of friends over time :** Here they try to find the rate of growth of friends of a real user and try to find similarity with the target account. They proposed a statistical model to find the threshold value of the growth rate.

2. **Social interaction :** If a fake account is smart enough and successfully mimic the growth rate of a real account then this property can help to detect the fake account. In here they check for the social activities of the target account with the victims account (friend of the target account). They assumed that it is very unlikely that a fake account will take part in a social interaction with it's victim because because for that the fake accounts owner would have to have sufficient knowledge about the victim's life.

3. **Evolution of the OSN graph over time :** Finally if the fake account knows its victim in person and managed to bypass the above two detection methods,then as the last filter they check the evolution of the OSN graph over time. Here they try to find similarity between the target account and a typical account with respect to two factor:

   (a) Average degree of the nodes in the OSN graph

   (b) Number of singleton friends in the OSN graph

   Using statistical method they evaluate the mentioned factors and try to predict the validity of the fake account.

Chapter 4

Data Sets

For our research we needed real world facebook user data. Thought some social graph datasets are available publicly but those does not fulfill our requirement thus we had to create our own datasets. We collected user information from our own accounts as well as from our friends accounts. Detail description of the data collection procedure is mentioned in the following section.

4.1 Data Collection Procedure

Facebook has it's own official API for providing user information but is very restrictive due to privacy issue. As we needed a user's activity and behavioral based data to differentiate between fake and real users Facebook's API were not sufficient. So we build our own automated tool that collects necessary information without violating any privacy policy.

The tool automatically logs into Facebook taking user credentials as input and then browse to the profile of each friend the user has, from there it collects the required information and creates a JSON object containing the informations. Lastly the JSON is saved into a file. Fields our tool take into concern are listed below

1. About
2. Basic-info
3. Contact-info
4. Education

5. Family

6. Living

7. Number of photos

8. Number of post tagged in

9. Number of total post by user

4.2 Data Moderation and Labeling

We collected 1310 user data among which 1210 are labeled as real user and 100 are labeled as fake user. For labeling a user as fake or real we visited profile of each user and make sure we are absolutely sure about the user being real or fake. Most of the real users were selected based on real life interaction with the authors so that their remains no doubt about the validity of the users being real. For fake accounts we visited the user profile very carefully to find anything that can verify the user as real, we also contacted the users through facebook messages option to be sure about our labeling. Anytime we face a dilemma about a user being fake or real we left out the user form our dataset.

Chapter 5

Methodology

In this chapter we shall discuss about our own approach towards detecting a fake account. We have divided the chapter into two parts namely **Feature Selection** and **Methodology**. In the Feature Selection section feature identification and justification have been discussed and in Methodology section we have discussed about our proposed model.

5.1 Feature Selection

In our data collection process, we crawled Facebook user data which are at least tagged as 'Only Friend'. We use 3 honeypot account to collect these data. In the feature selection procedure, we give priority in 3 feature as well as we used 13 feature as total. Our main observation was

1. A fake user doesn't upload photos more frequent, in many cases they use less than 10 photo in their lifetime.

2. A fake user is less likely to tagged in photos than other users. Although this assumption is not valid in Celebrity Profile. But at this moment Facebook used 'Blue Tick' to recognize the valid account of public figures, and the authority are strict on fake public figure account, so it is considered as a powerful feature in our experiment.

3. We considered the the percentage of tagged photos over last 100 post of the user.

Full feature list with overall description is given in Table 5.1. Here, we do not considered the whole Birth Day of a user, rather we used Birth Year, because from our observation we saw that,

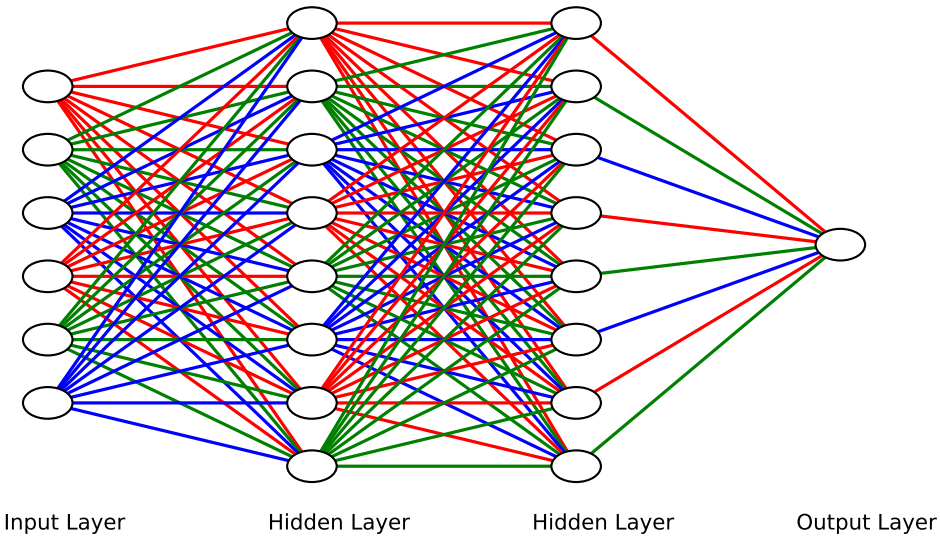

Figure 5.1: Deep Neural Network

most fake account hide their birth year in their profile, a majority also uses '1905' as the Birth Year because it is Facebook default when an newcomer wants to open a Facebook account.

5.2 Experiment

Our main challenge on this work was to select a classifier with good accuracy. Previously many authors used Naive Bayes, KNN, SVM, Decision Tree and many others. We choose Deep Neural Network (DNN) as our primary classifier. As we have a limited number of data sets available we divided the whole experiment into two segment. In first segment we used the standard ratio of 60:40 and in second segment of the experiment the Train,Test ratio was 70:30.

Table 5.1: Feature Set Table with Description and Justifications

Feature Name	**Description**	**Justification**	**Measuring Method**
gender	Gender of the user.	Probability of a fake account created as female account is high.	Gender of a account can be obtained from user profile.
birthday	Birthday is displayed or hidden.	Fake accounts tends to hide birthday.	Birthday can be obtained from user profile.
mobile	Mobile number is displayed or hidden.	Fake accounts are expected to hide mobile number.	Mobile number can be obtained from user profile.
instagram	If user has linked Instagram account.	Real users are expected to have Instagram linked with Facebook.	Instagram account can be obtained from user profile.
home_town	if home town is displayed.	Fake users are expected to hide home town.	Home town can be obtained from user profile.
current_city	if corrent city is displayed.	Fake users are expected to hide current city.	Current city can be obtained from user profile.
education	In how many institution the user has attend.	Fake users are expected to hide institution where and when they attended to.	Education can be obtained from user profile.
no_of_work	In how many places the user have worked.	Fake users are expected to hide institution where and when they have worked	Works can be obtained from user profile.
no_of_family	Number of family member linked.	Real users tends to have more family member linked.	Family can be obtained from user profile.
photo_count	Number of posted photos.	Real users are expected to have more photos.	Number of post can be count from user's timeline.
tagged_in	Number of post the user is tagged in.	Real users are expected to be tagged in more post.	Number of post in which the user is tagged in can be count from post in user's timeline.
total_post	Number of post the user has published.	Real users are expected to have more original post.	Total number of post the user has published can be found on user's timeline.

Chapter 6

Result and analysis

6.1 Result

As mentioned earlier based on training and test data ratio our experiment was divided into two phase. Standard Train and Test data ratio is 60:40 but we assumed that as we have a small number of data we shod follow a custom approach. So we divided the data into 70:30 ratio and carried out the same experiment. The result and comparison of both the experiment is given below.

6.1.1 Standard Ratio (60:40)

In this experiment we trained the model with 786 user data And evaluated the with over 524 user data. In which our model marked 18 fake user as fake and 477 Real user as a real user. The confusion matrix is shown in Table 6.1 and the accuracy graph over steps is shown in Figure 6.2

Summary of this experiment can be expressed as

- Precision: 72.00%
- Sensitivity: 45.00%
- Accuracy: 94.47%
- Specificity: 98.55%

Table 6.1: Confusion Matrix (Standard Ratio)

n = 524		Output	
		Fake	Real
Expected	Fake	18	22
	Real	7	477

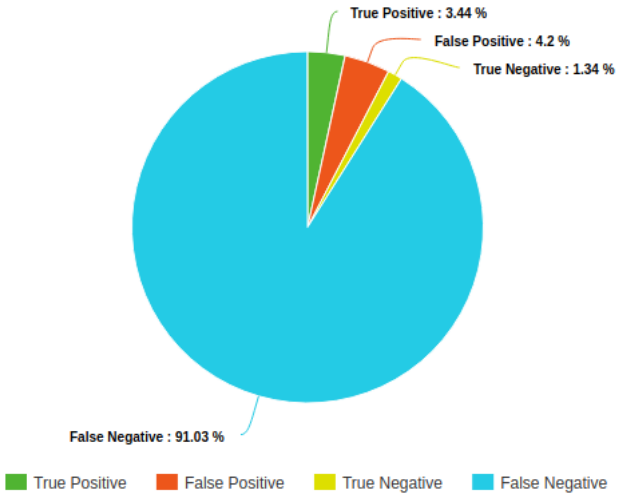

Figure 6.1: Standard Ratio Result Summary

6.1.2 Custom Ratio (70:30)

In this experiment we trained the model with 916 user data. And evaluate the model over 394 user data. In which our model marked 15 fake user as fake and 361 Real user as a real user. The confusion matrix is shown in Table 6.2 and the accuracy graph in shown in Figure 6.4

Summary of this experiment can be expressed as

- Precision: 88.24%
- Sensitivity: 48.39%
- Accuracy: 95.43%
- Specificity: 99.45%

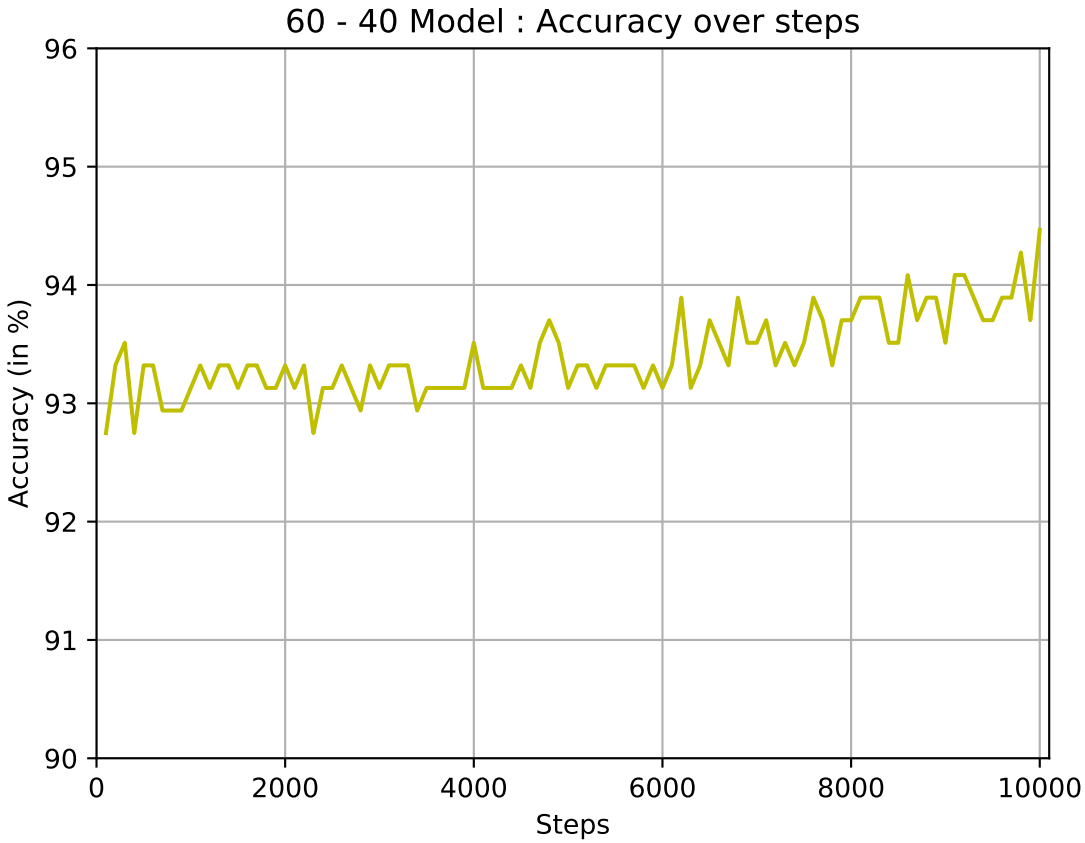

Figure 6.2: Accuracy over step using Standard Ratio(60:40)

Table 6.2: Confusion Matrix (70:30 ratio)

n = 394		Output	
		Fake	Real
Expected	Fake	15	16
	Real	2	361

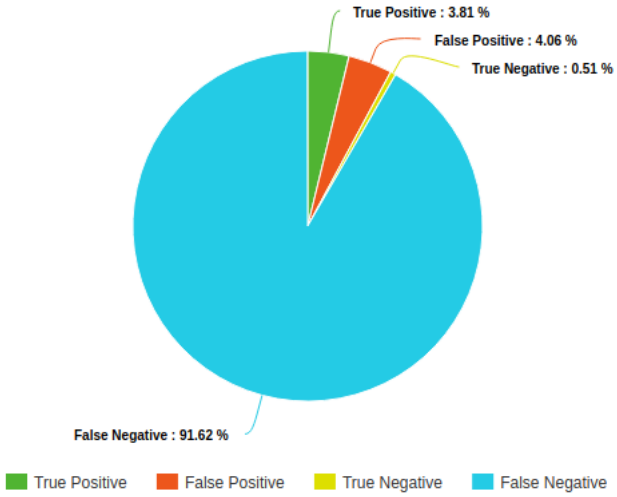

Figure 6.3: Custom Ratio Result Summary

6.2 Analysis

As shown in Figure 6.5 we can compare the performance of our model in the nature of dataset. In both experiment we got accuracy over 90%. But we've got poor result on Sensitivity. The main reason behind this is, the ratio of output class is only 92:8 where 92% is labeled as Real user and rest 8% is labeled as Fake user. So it is quite tough for our model to determine expected output.

Next, when we switched from 60-40 ratio to 70-30 ratio, resulted sensitivity rise about 3%. Table 6.1 & Table 6.2 show us the details behaviors of the training. In the second experiment the model is much better in identifying real and fake user. That's why False Positive is lesser in the second experiment.

6.3 Comparison

Most of the previous works done in this topic has a high accuracy [10, 12]. We could not get the datasets they used for their experiment. But reviewing their approach we think they used a biased datasets to get high accuracy.

At this time we also have a poor dataset so verifying their models with our own dataset would not help much. Thats why we decided to do the comparison when we will enrich our dataset.

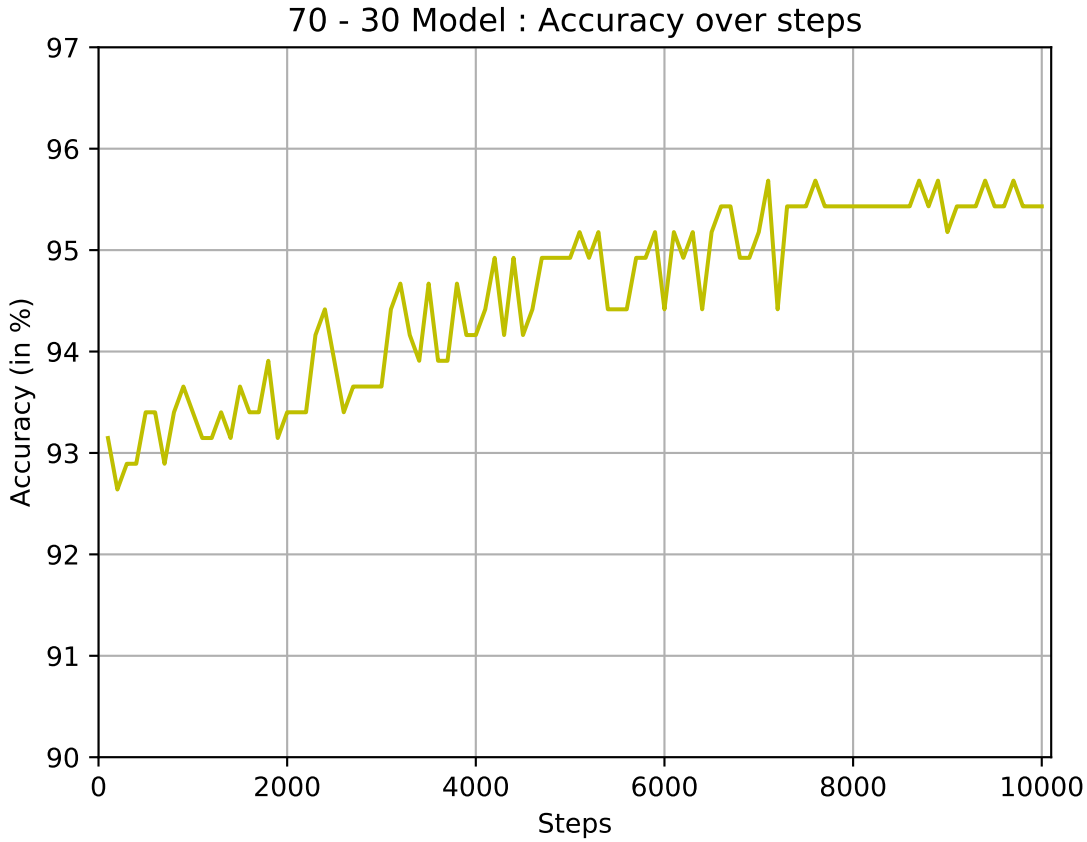

Figure 6.4: Accuracy over step using Custom Ratio(70:30)

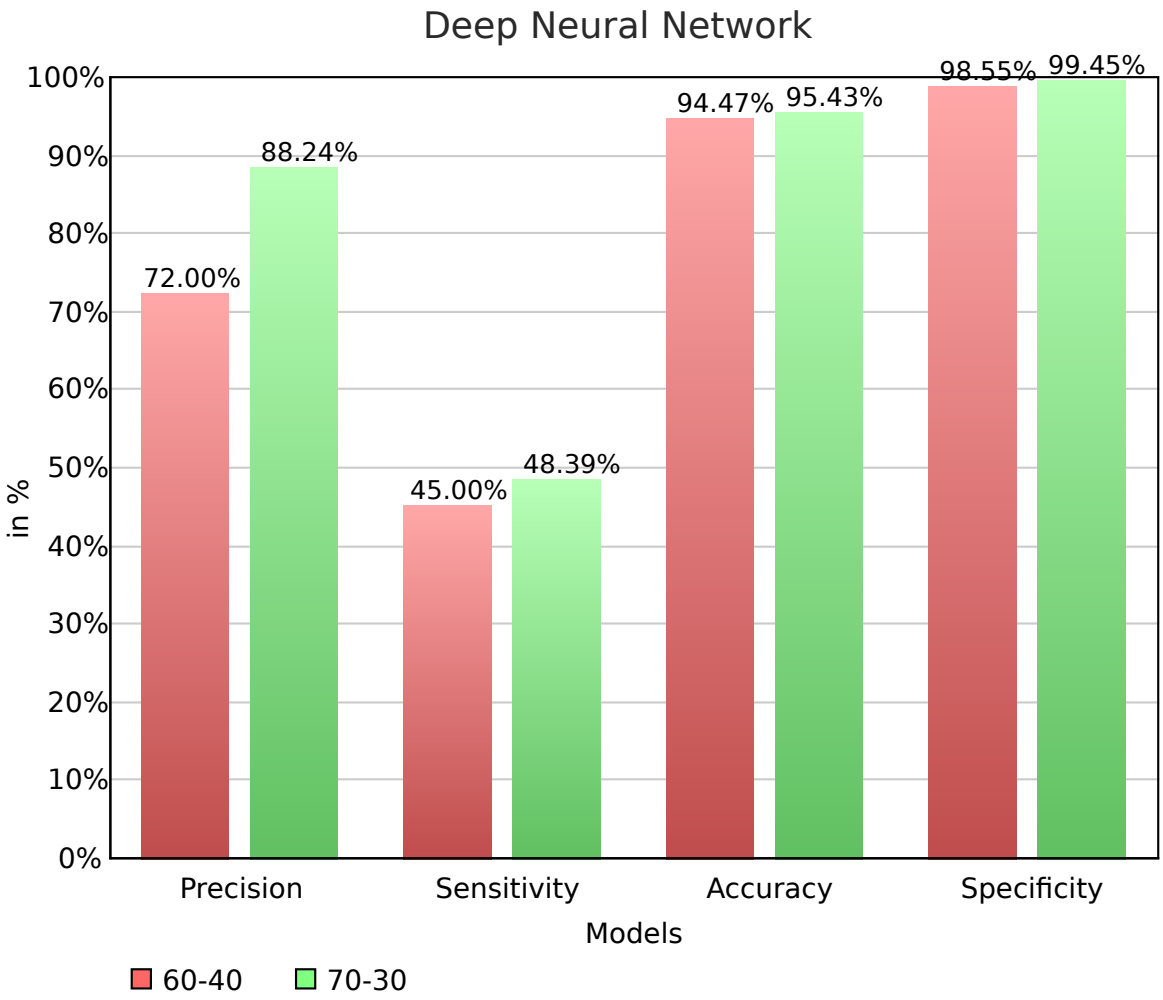

Figure 6.5: Comparison between 60:40 and 70:30 data ratio

Chapter 7

Conclusion

From our experiment we identified some major plus points and also some drawbacks of our approach. In the following section we shall discuss about some of them and also determine our future works.

7.1 Discussion

We collected our data from facebook and as mentioned earlier facebook has a very restricted privacy policy. As the API facebook provide is not very useful we had to build our own tool to collect our data. The data collection procedure was very complex and thus we could not get as much data as we needed. Also data filtering and labeling was a very time consuming process. Facebook has changed rapidly over the past five years thus there is a very huge difference between todays users and users from five years ago. Thats why feature selection can not be same for both types of user and it was one of the main reason our method did not perform up to our expectation. Also we think the set of selected feature was not sufficient in many cases. But our overall accuracy was good enough and with some small change and more data our method can perform well.

7.2 Future Work

From the result of the carried out experiment we can see that our accuracy is promising, but at the same time the sensitivity is very poor. The drawbacks were not hard to find out. If we take a

closer look to our data sets and our proposed model we can clearly see some important points that needs to be improved. In this section we shall discuss about those.

7.2.1 Improved crawler

Our custom crawler can collect user data automatically but is limited to a number of fields. As for now it only collected data we are using in our work, but if we could build a improved tool which will be able to collect more information from a profile then the features will be much more precise as we will then have more data to work with. Currently we are not taking into account many fields simply due to the fact that it will take a huge amount of time to collect data from those fields. With a improved and more powerful crawler we will be able to overcome this and our accuracy of the model will improve too.

7.2.2 New Feature Set

Currently we have selected 12 feature based on which our model will classify the data sets. Although these features are important but not sufficient. There exist many cases to look closely. For the last five years the user and Facebook both have changed so much. We need new feature set that will be good for both past and present user. Also currently we have overlooked some feature because data collection about those feature is too much complicated. If we can simplify these fields then we can add more features in our feature list.

7.2.3 More data

One of the most important reason behind our poor sensitivity was that we did not have sufficient data to train our model more accurately. The number of Facebook user is not less but for collecting data that we need we have to pass certain privacy policy that is to get the data we need we must have to be friend of that user. It is very difficult to collect data in this scenario. Also data labeling is one of the most challenging part of our data processing phase. Our custom crawler collects user data automatically but not smart enough to label them. So we have to label them manually and this takes a lot of time. But if we could collect more data specially more fake user's data then it would be very helpful for our model. With a large amount of data we are certain that our proposed model will give a more improved accuracy and sensitivity.

7.2.4 Verify Previous Model

Due to poor dataset we could not verify or compare our model with previous model. Though many previous authors claimed that their models has very high accuracy but we think with an unbiased datasets they would give different result. After improving our crawler and collecting new and more datasets we want to verify all the existing model against our dataset. Finally we will compare their results with our result.

References

[1] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Towards detecting anomalous user behavior in online social networks," in *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, 2014, pp. 223–238. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/viswanath

[2] R. G. Michael Fire and Y. Elovici, "Online social networks: threats and solutions," *IEEE Commun Surv Tut, 16 (4) (2014)*, pp. 2019–2036, 2012.

[3] M. Fire, G. Katz, and Y. Elovici, "Strangers intrusion detection - detecting spammers and fake profiles in social networks based on topology anomalies," vol. 1, 01 2012.

[4] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, "Advanced social engineering attacks," *Journal of Information Security and Applications*, vol. 22, pp. 113–122, June 2015. [Online]. Available: https://doi.org/10.1016/j.jisa.2014.09.005

[5] T. Stein, E. Chen, and K. Mangla, "Facebook immune system," pp. 8:1–8:8, 2011. [Online]. Available: http://doi.acm.org/10.1145/1989656.1989664

[6] V. Sharma, S. Rai, and A. Dev, "A comprehensive study of artificial neural networks," 10 2012.

[7] S. Gurajala, J. White, B. Hudson, and J. Matthews, "Fake twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," vol. 2015, 07 2015.

[8] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in *Proceedings of the 2015 International Conference on Social Media & Society*, ser. SMSociety '15. New York, NY, USA: ACM, 2015, pp. 9:1–9:7. [Online]. Available: http://doi.acm.org/10.1145/2789187.2789206

[9] S. Adikari and K. Dutta, "Identifying fake profiles in linkedin," *Pacific Asia Conference on Information Systems*, 2014.

[10] A. Gupta and R. Kaushal, "Towards detecting fake user accounts in facebook," *2017 ISEA Asia Security and Privacy (ISEASP)*, pp. 1–6, 2017.

[11] M. Conti, R. Poovendran, and M. Secchiero, "Fakebook: Detecting fake profiles in on-line social networks," pp. 1071–1078, 08 2012.

[12] S. Y. Wani, M. Kirmani, and S. Imamul Ansarulla, "Prediction of fake profiles on facebook using supervised machine learning techniques-a theoretical model," vol. 7 (4), pp. 1735–1738, 08 2016.