A HYBRID APPROACH FOR CREDIBILITY DETECTION IN TWITTER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ALPER GÜN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

FEBRUARY 2014

Approval of the thesis:

A HYBRID APPROACH FOR CREDIBILITY DETECTION IN TWITTER

submitted by **ALPER GÜN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Assoc. Prof. Dr. Pınar Karagöz _____
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Ahmet Coşar
Computer Engineering Department, METU _____

Assist. Prof. Dr. Aybar C. Acar
Graduate School of Informatics, METU _____

Assist. Prof. Dr. İsmail Sengör Altıngövde
Computer Engineering Department, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ALPER GÜN

Signature :

ABSTRACT

A HYBRID APPROACH FOR CREDIBILITY DETECTION IN TWITTER

Gün, Alper
M.S., Department of Computer Engineering
Supervisor : Assoc. Prof. Dr. Pınar Karagöz

February 2014, 64 pages

Nowadays, microblogging services are seen as a source of information. It brings us a question. Can we trust information in a microblogging service? In this thesis, we focus on one of the popular microblogging service, Twitter, and try to answer which information in Twitter is credible. Newsworthiness, importance and correctness are the dimensions to be measured in this study. We propose a hybrid credibility analysis which combines feature based and graph based approaches. Our model is based on three types of structures, which are tweet, user and topic. Initially, we use feature based learning to construct a prediction model. In the second step, we use the results of this model as input to authority transfer and further refine the credibility scores for each type of node. The same process is used for measuring each of the dimensions of newsworthiness, importance and correctness. Experiment results show that the proposed hybrid method improves the prediction accuracy for each of these credibility dimensions.

Keywords: Credibility, Twitter, Microblog, Random Forest Decision Tree, Authority Transfer

ÖZ

TWİTTER'DA GÜVENİLİRLİK TESPİTİ İÇİN BİR HİBRİT YAKLAŞIM

Gün, Alper
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Pınar Karagöz

Şubat 2014 , 64 sayfa

Mikro ağ güncelerinin günümüzde bilgi kanağı olarak kullanılması beraberinde cevaplanması gereken bir soruyu getiriyor. Microblog sitelerinde bilgilere güvenebilir miyiz? Bu çalışmada popüler microblog sitelerinden birisi olan Twitter'a odaklanıyoruz ve Twitter'da hangi bilginin güvenilir olduğunu cevaplamaya çalışıyoruz. Tweet'leri haber niteliği taşıyıp taşımaması yönünden, önemli olup olmaması yönünden ve doğru bilgi içerip içermemesi yönünden değerlendiriyoruz. Bu tezde çizge ve özellik temelli çalışmalardan temel alarak hibrit bir sistem sunuyoruz. Modelimizde tweet, kullanıcı ve konu olmak üzere üç tip yapı bulunmaktadır. Öncelikle özellik temelli öğrenme yöntemini kullanarak bir model kuruyoruz. İkinci aşamada ise bu modelin sonuçlarını her düğüm tipi için otorite transferine girdi olarak veriyoruz. Aynı süreci haber değeri taşıması, önemli olması ve doğru bilgi taşıması yönünden tekrarlıyoruz. Deney sonuçlarına gore önerdiğimiz hibrit yöntem her üç güvenirlilik etiketinde de doğruluğu artırmaktadır.

Anahtar Kelimeler: Güvenirlilik, Twitter, Mikro Ağ günceleri, Haber, Random Forest
Karar Ağacı, Otorite Transferi

To my family and people who are reading this page

ACKNOWLEDGMENTS

First of all, I am grateful to my supervisor Assoc. Prof. Dr. Pınar Karagöz for her guidance and support during my thesis. Being her student is a pleasure and privilege for me. Once again, I would like to thank her.

I would like to thank to Prof. Dr. İsmail Hakkı Toroslu, Assoc. Prof. Dr. Ahmet Coşar, Assist. Prof. Dr. Aybar Acar and Assist. Prof. Dr. İsmail Sengör Altıngövde for their kindness to accept participating in the jury and spending their valuable time for evaluating my thesis.

I would like to express my gratitude to all my teachers in my life for contributing my success so far.

I am also grateful to my dear friend Özge Uyanık for her support and motivation.

Last but not least, special thanks to my parents and brother for their endless love and support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

GT_{avg}	Ground Truth for Average Evaluation
GT_{4YES}	Ground Truth for 4 YES Evaluation
GT_{3YES}	Ground Truth for 3 YES Evaluation
GT_{1YES}	Ground Truth for 1 YES Evaluation

CHAPTER 1

INTRODUCTION

Microblogging services are used by many people to share contents such as news, comments, images or videos. The difference of microblogs and traditional blogs is the size of content. Some of the microblogging services allow their users to share their comments in limited characters and some of them limit the size or duration of media files. Generally they have a friendship mechanism and each user broadcasts his/her post to other people. As more people use a microblog service, service reaches more people by using friendship network of users. There exits many microblogging services but we can list some of the popular ones such as Twitter [30], Facebook [8], Tumblr [29], Yammer [39], Instagram [14] and Vine [33].

Twitter is one of the most popular microblogging and social networking services which is used worldwide by millions of people. In Twitter, posts of users are named as ***tweet*** and each tweet may contain at most 140 characters. For this reason, users need to express themselves with less number of words to keep the content with 140 characters. If they need to share a URL, they may need to use a URL shortening service. These links can contain URL to images or videos in other websites and users can view these media content inside Twitter application.

Friendship in Twitter is not mutual. Users can follow any other users and be aware of tweets of these people. There is no need to mutual friendship and this flexibility makes Twitter easy to get information and news for a topic or a person. Twitter accounts are generally owned by real users or corporate organizations. But there may exist some fake accounts. Therefore Twitter has verified user flag for accounts, which indicates that the account owner is real. There is a special sign **#** which means

hashtag in tweets. Hashtag provides to group relevant tweets under the same topic. Users can comment on a specific topic by using hashtags. Topics which have most tweets in a specific interval are announced as a trend topic in Twitter. Most popular 10 topics are listed in Twitter for each region. Another important feature of Twitter is ***retweeting***. A user can retweet by using another user's tweet. He/she can add his/her own post to original tweet or he/she can exactly send original tweet. ***RT*** keyword is used in retweets to indicate that source of this tweet belongs to someone else. Besides, Twitter has one more special keyword which is @ sign. This sign is named as ***mention*** and users write it together with target user's name when referring to this user in their posts. Authors in Twitter mostly prefer to use acronyms and abbreviations since there is a constraint on maximum number of character in a tweet.

In Twitter, there exist many Twitter accounts that are followed by many people to get the latest news or comments. Almost all well-known news agencies have their own Twitter accounts and they publish their news also from their Twitter account. Furthermore, people follow ideas of many artists, politicians, journalists etc. There are many academic studies that use huge amount of data in Twitter. For example, one of the popular research topics on Twitter is recommendation systems. There are various studies for recommending links, news [2], information sources [4] etc. Furthermore, there are studies about detection of important events such as earthquake [26]. These studies analyze tweets to obtain a result. For example, earthquake detectors need to analyze all tweets which are posted in last few minutes and lead to a result about earthquake detection. However, these studies require picking proper ones among all tweets. For this reason, there is a need to investigate the credibility of users and tweets before using this massive amount of data. Some of tweets may contain incorrect information and it leads such systems to wrong decisions. Besides, some of users may be more valuable for their posts since their posts contain important information but some of them may have no impact on other user. Therefore there is a need to classify users and tweets on the basis of importance. After that, a decision system can rely on important users and tweets more than other users and tweets.

Credibility problem in microblogging services is studied in several studies. We can classify them into three groups. Studies in the first group build a machine learning model and aim to learn credibility value from the data [6]. They generally use at-

tributes of authors, posts and topics. The second group of studies measure credibility by utilizing friendship and retweeting network in Twitter [38]. A few studies also use topic relationship in this network. Algorithms such as PageRank [23] and HITS [17] are commonly used techniques to distribute score in the network. Studies in the third group are hybrid solutions. They use approaches in the first group and second group together to make a decision for credibility value in microblogging services.

In this thesis, we focus on classification of tweets and users in terms of their newsworthiness, importance and correctness. Our study can be used by any application which tries to find important, newsworthy and correct data among millions of tweets. We propose a hybrid approach and apply a two-level process for ranking tweets in terms of these dimensions. In the first step, we build a decision tree to classify tweets, users and topics with their attributes. We have in total 41 attributes for user, tweet and topic. We get the best score with random forest decision tree algorithm which has success rate about 80-90 percent. In the second step we apply authority transfer, which gets initial scores from the first step. We have three different types of nodes which are user, tweet and topic. There are undirected edges in our graph which are tweet-user edges and tweet-topic edges. Each tweet is linked to a user and a topic. It allows us to transfer authority among user, tweet and topic nodes. The authority transfer makes use of the idea that if a user is important then we can conclude that tweets which are posted by this user are also important. Similarly if a topic is important, tweets in this topic are also important. Each node starts with initial score coming from feature evaluation. Besides, number of followers is added to the initial score to user nodes and number of retweets is added to the initial score to tweet nodes. After we transfer score between nodes, we obtain a final score for each node and if score of a node is more than a predefined threshold, then we label this node as newsworthy, important or correct. Otherwise we label this tweet not newsworthy, unimportant or incorrect. We measured the success of our method by in comparision to user annotations for newsworthiness, correctness and importance of a tweet. After authority transfer, we measured about 91% success rate when predicting newsworthiness, 86% percent success rate when predicting importance and 84% percent success rates when predicting correctness.

Contribution of this study is as follows.

- Our study takes advantage of feature based and graph based approach. Most of the studies in the literature use one of the approaches. There exist some hybrid approaches but they only uses friendship network to transfer authority.
- One of our based study uses authority transfer among user and tweets. [38] We add topic as a third node to this graph structure since credible topics may contain more credible tweets and incredible topics may contain less credible tweets.
- We measured credibility in terms of three criteria which are newsworthiness, importance and correctness. However, most of the studies in the literature focus one of these criteria.
- We applied our thesis into Turkish tweets. It is also possible to use our system to other languages. Only language dependent part in this study is sentiment analysis and personal pronouns.

This thesis is organized as follows. Chapter 2 gives a summary of research in the literature about ranking and credibility in microblogs. Chapter 3 discusses the detail of our proposed solution. Building our data set, user evaluation and proposed work are mentioned in this chapter. Chapter 4 explains experiment results of proposed solution and finally Chapter 5 gives a summary of thesis and possible improvements in the future.

CHAPTER 2

RELATED WORK

There are two common ways to rank tweets, which are feature based evaluation and graph based evaluation. There are also hybrid approaches that combine feature based and graph based solutions. In this chapter, we summarize the research in the literature for each of these approaches.

2.1 Feature Based Solutions

Feature based solutions generally aim to build a learning scheme such as decision tree, neural network, SVM or Bayes network. They may use attributes of users, context and behavior.

Research of Castillo et al. [6] is in this category and it is one of the basic studies which inspire us. In [6], the aim is to classify tweets as credible and not credible. Castillo et al. use wide range features that are grouped as message based, user based, topic based and propagation based. Most of the features in our study are also derived from this work. Message based features contain structural attributes of tweet. User based features are related with the account detail of author in Twitter. They also categorize tweets into different topics. Each topic contains its own tweets. Topic based features are also calculated by getting the average of message based and user based features. Further, they use propagation based features which are related with retweet tree. Twitter Monitor [20], which is an application to detect important events in Twitter is used in their research and they extract tweets during two months. After data is collected, statistical study is done to classify tweets as newsworthy or not.

They used Mechanical Turk [1] which provides functionality to get feedback of many users about prepared questions. User feedbacks and tweet data set are trained to automatically find credible tweets. Their study has precision and recall rate between 70% and 80%.

There are various studies which uses learning schemas. O'Donovan et al. [22] use features and focuses on credibility assessment by using features such as URLs, mentions, retweets and tweet length. Jenders et al. In [15], the authors investigate what features makes a tweet viral. This study utilizes two types of attributes, which are obvious features and latent features. Obvious features in this study contains user and message based features such as tweet length, number of followers etc. Latent features contain emotional attributes and sentiment analysis. Another study also uses content based, tweet based, user based features and rank by significance of tweets [32]. The study of Pal et al. [24] classifies users by examining the features of each tweet and user profile. They rank users in a given topic in terms of their authority. They categorize tweets into three groups, which are original tweets, conversational tweets and repeated tweets. Conversational tweets are pointed to another user by using mention tag. Repeated tweets share originally written tweet by retweeting it. Original tweets are written by author itself which are not in the category of conversational and repeated tweets. Alonso et al. [3] investigate effectiveness of 13 features, which are mostly content based, and then label tweets such as interesting, important or spam. Another research also investigates the spam issue[42]. This study aims to detect spam and promotional campaigns. They classify messages in Twitter as regular messages, promotional messages and spam messages. In order to detect spam and promotional messages, they analyze similarity of URLs by utilizing tweet based attributes. CredRank [1] algorithm measures user similarity by clustering user's behavior. They measure credibility not only for Twitter but also for other social media platforms. Credibility of information in blog sites is also investigated in [34]. This study focuses on the content and checks some credibility indicators such as spelling, timeliness, document length and comments. They propose a ranking strategy for blogs in terms of their credibility and select top n blogs as credible.

Trustworthiness of tweets is studied in several studies by using learning schemas.

[1] https://www.mturk.com/mturk/

They examine key elements of each post and help users to evaluate trustworthiness of a tweet [21]. In [10], Gupta et al. use message based features and user based features to acquire trustworthy tweets in high impact events. They investigate tweets during 14 important events such as Libya crisis, hurricane Irene, earthquake in Virginia and UK Riots. Since all tweets are not trustworthy, their study provides an analysis on quality of information in Twitter. Xia et al. [37] aim to measure trustworthiness of tweets in emergency situation and labels tweets as credible or incredible by using bayesian network. This study analyses author based features, content based features, topic based features and diffusion based features.

Wikipedia is also used to calculate credibility in [28]. They investigate user posts in social network services such as LinkendIn and Facebook. They compare message with article in Wikipedia. If similarity is strong, they conclude that user post is credible.

Yang et al. try to answer effect of cultural differences on credibility. [40] They focused on two microblogs which are Twitter for United States and Weibo [35] for China. They investigate the impact of some features on credibility perception among people from United States and China. Their feature list contains author gender, author name style, profile image, location and friendship network.

2.2 Graph Based Solutions

The second common way for credibility ranking on Twitter is graph based solutions. Studies in this group build a graph with nodes of user, tweet and topic and then transfer score between nodes.

While transferring authority between nodes, in the literature there are two common algorithms which are PageRank [23] algorithm and HITS[17] algorithm. PageRank analyses distribution of links and calculates a probability score to access randomly a web page. Each node shares its score to other nodes iteratively. Similarly HITS algorithm also distributes score of nodes to each other. As the basic difference, it calculates hub and authority scores and transfers them to other nodes. Hub score denotes the value of neighbor nodes and authority score denotes the node itself. It

is not effective to use page ranking and HITS algorithm for evenly distributed data sets such as each node has same number of neighbors. Since PageRank and HITS algorithms calculate the same score for these kinds of data sets, we cannot apply these ranking algorithms in symmetric data sets.

TURank, which is one of the base studies for our research, uses actual information flow in Twitter to find authoritative users [38]. They apply a few object ranking algorithms and transfer authority between tweets and users. According to TURank, a user is more authoritative if this user is followed by an authoritative user. Similarly if a tweet is retweeted by an important tweet, it makes the original tweet more important. A sample authority transfer graph seen in Figure 2.1.

Figure 2.1: Sample Graph Used in TURank. Retrieved from [38]

When they transfer authority, they use object ranking which is an extension of PageRank algorithm. There are four ways to transfer authority in TURank:

- User to User: A user transfers his/her score to his/her followers.
- User to Tweet: A user transfers his/her score to his/her all tweets.
- Tweet to User: A tweet transfers its score to its author.
- Tweet to Tweet: If a tweet is retweeted, retweet transfer its score to the original tweet.

They try four different sets of weights for edges in TURank and compare their results to HITS and Page Rank algorithms.

User followership network is used by many studies to rank users in Twitter. The work of Armentano et al. [4] aims to recommend interesting users to follow. They utilize user's follower graph to improve recommendation quality. The same authors also

perform another research which uses a user topology for recommending good information sources [5]. Another research focuses on retweet tree and finds interesting tweets [41].

There are several studies which that also use topic as a node in their graph structure. Kong et al. [19] try to rank users, tweets and topics by using topic focus degree, retweeting behavior and influence of users. They use weighted directed graph when transferring authority. Similar to this research, another paper also observes followership, retweet and mention trees [7]. They calculate influence score of users by using these trees across topics and time. Gupta et al. [11] use events instead of topics. They try to evaluate reliable events by using a PageRank like algorithm and three layers of graph, consisting of user, tweet and event. They claim that their method gives more accurate results than classifier based solutions.

Some studies use web page links as another type of node. In [25], authors model a graph with three nodes which are user, tweet and web pages. Scores of each node are aggregated with their relationships to other node. Another work, Tri-HIT [12], provides a tweet ranking algorithm that works in web - tweet - user heterogeneous networks.

2.3 Hybrid Solutions

Hybrid solutions which uses feature based solution and graph based solution also exist in the literature. Kang et al. [16] provide a hybrid model solution and calculate a score by using 19 features in total and propagate this score in their network. They only use user friendship network and users sends their score to their followers. Another hybrid solution is proposed in [9]. This study ranks tweets in order to find the most retweetable posts. They use a hybrid model which uses small set of features consisting of user, publisher, tweet, user - publisher and user - tweet. These feature scores are transferred to other nodes in their graph. The study of Huang et al. [13] proposes a hybrid strategy to calculate influence of users. They uses page ranking algorithm in user friendship network. Further, they utilize user behavioral attributes such as frequency of updating microblog, interaction with other users, and so on.

We will also use a hybrid solution in this study. However, hybrid solutions in the literature utilize content based and graph based solutions limitedly. For example, they do not use sufficient attributes for learning schemas. Besides, they generally focus on user friendship network. However, in our study we use much more number of attributes for our learning schema and our graph structure contains not only user relationship but also tweet and topic relationships.

CHAPTER 3

PROPOSED SOLUTION

There exist many studies which concentrate on microblog credibility problem in the literature. Some of these studies aim to train and learn data and then predict credibility of tweets. Moreover, some of the studies in this area go with solution which utilize relationship of users, tweets etc.

Studies of Castillo [6] and Yamaguchi [38] form the basis of our hybrid solution. We start the training phase with machine learning schemes and then use authority transfer on our graph, which consists of user, tweet and topic relationship. Yamaguchi's study involves a graph with user - tweet relationship. In our thesis, we also use topic - tweet relationship because it is possible that some tweets may have more credible if this tweet is in a credible topic. Therefore we also decided to transfer authority between tweet and topic. Solutions in the literature generally measure the credibility in terms of a single labels. However, we use three labels which are newsworthiness, importance and correctness, and measure credibility in each of these dimensions.

Figure 3.1 represents flow in our solution. We firstly collected data by using Twitter API [31]. The proposed method basically builds a prediction model by using a learning scheme and then these prediction results are used in authority transfer. In order to build a prediction model, we need to collect data and construct a set of tweets annotated for newsworthiness, importance and correctness. In the rest of this chapter, we present the details of data collection and the user study as well as the proposed technique.

Section 3.1 gives information about building data set. User study is explained in

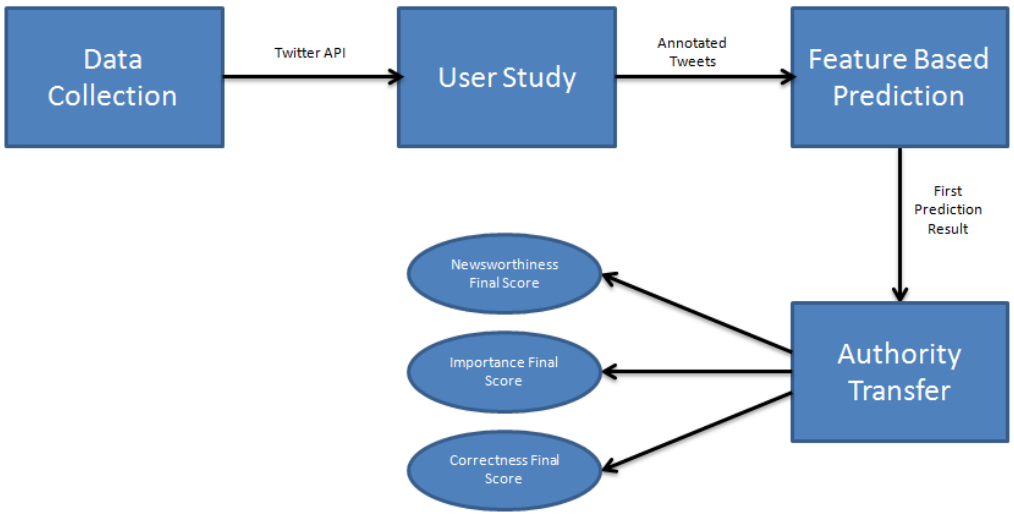

Figure 3.1: Flow Diagram of Our Proposed System

Section 3.2. Feature based study is given in Section 3.3 and graph based study is discussed in Section 3.4.

3.1 Data Collection

Since we use topic in our solution, we want to have equal number of tweets for each of topic in our data set. Otherwise, different number of tweets for each topic may cause unfair score in authority transfer. Therefore, we firstly determined the topics which will be used in our solution and then collected equal number of tweets for each topic

As the first step of data collection, 25 trend topics are determined. These are among the trend topics in Turkey which are announced by Twitter. Trend topics are selected in several time intervals between January 2013 and June 2013. List of chosen trend topics are presented in Table 3.1.

Table 3.1: Trend topics used in our data set

Topic Name	**Start Time**	**End Time**	**Number of Tweets**
Toktamış Ateş	19 Jan 2013	10 Mar 2013	100
Mehmet Ali Birand	28 Feb 2013	10 Mar 2013	100
Haydarpaşa	01 Mar 2013	10 Mar 2013	100
Galatasaray Üniversitesi	24 Feb 2013	08 Mar 2013	100
Fatih Ataylı	27 Feb 2013	10 Mar 2013	100
Danıştay	07 Mar 2013	10 Mar 2013	100
Barboros Şansal	10 Mar 2013	11 Mar 2013	100
Ankara	02 Mar 2013	10 Mar 2013	100
Adnan Oktar	01 Mar 2013	02 Mar 2013	100
Cumartesi Demek	19 Jan 2013	10 Mar 2013	100
Zeki Kayahan	25 May 2013	26 May 2013	100
Üstad Necip Fazıl	25 May 2013	26 May 2013	100
Türkiye	25 May 2013	26 May 2013	100
Tturenc	25 May 2013	26 May 2013	100
KeremCem	25 May 2013	26 May 2013	100
İstanbul	25 May 2013	26 May 2013	100
İibf	25 May 2013	26 May 2013	100
Galatasaray	25 May 2013	26 May 2013	100
Fenerbahçe	25 May 2013	26 May 2013	100
Cbabdullahgül	25 May 2013	26 May 2013	100
Ahmethc	25 May 2013	26 May 2013	100
Utkuali	25 May 2013	26 May 2013	100
SeniçkiyiYasaklıyorsunda	25 May 2013	26 May 2013	100
BenceYasaklansın	25 May 2013	26 May 2013	100
BenAnlamam	25 May 2013	26 May 2013	100

Twitter API [31] enables users to extract public data from Twitter but it does not allow accessing some user profiles because of privacy concern. Therefore, we only accessed

tweets posted by users with public profile. There is a restriction to use Twitter API. It is possible to send 180 get requests in each 15 minutes interval.

We collected 100 tweets for each topic and in total we collected 2500 tweets for 25 topics. Twitter API services used to get the latest 100 tweets in each topic. We worked on tweets written in Turkish but we can also apply the same process to other languages.

After we extracted tweets from Twitter API, we parsed tweets to calculate tweet features. For features of user, we again used Twitter API and extracted user information and added this information to our data set. For topic features, we calculated averages of user and tweet features within a topic. There are 100 tweets in each topic and we calculated topic features by getting the average of 100 tweets and 100 users.

3.2 User Evaluation

To construct the ground truth, each tweet in our data set has been evaluated by four users. There are total 2500 tweets and for each tweet in data set, three questions are asked to each evaluator. Question and answers are listed in Table 3.2.

Table 3.2: Questions and answers for user study

Question	**Answer**
Is Newsworthy	YES NO NEUTRAL
Is Important	YES NO NEUTRAL
Is Correct	YES NO NEUTRAL

Newsworthy: We asked user to select newsworthy, if a tweet contains information which is important enough to report as news.
Important: We asked evaluators to select important, if a tweet has important information for evaluator.
Correct: We asked evaluators to select correct, if a tweet seems to have true information for evaluator.

We designed an application with user interface to help users for their evaluation as seen in Figure 3.2. In this application, users read each tweet one by one and mark the tweet for three criteria which are newsworthiness, importance and correctness. After users answer these three criteria for a tweet, we save their evaluation in a output file.

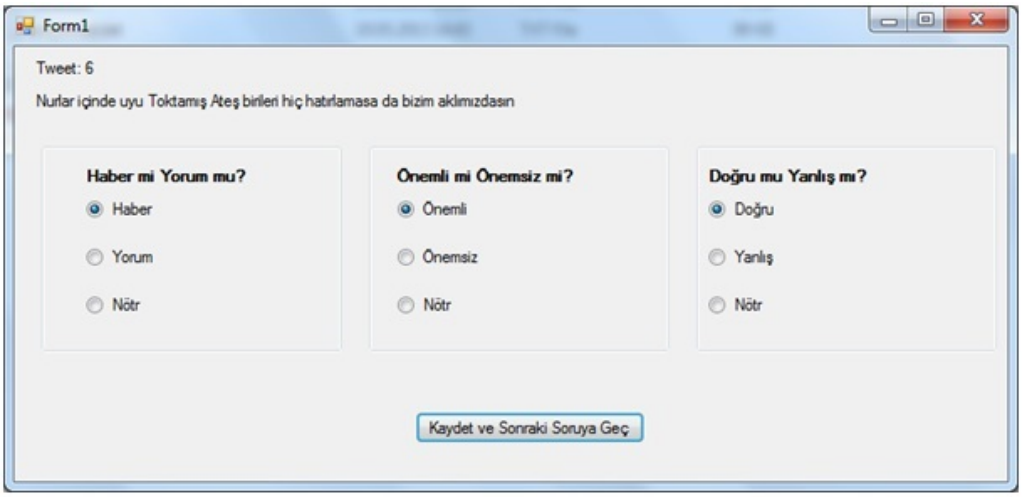

Figure 3.2: Evaluation Form

Respond correlations for each criterion are given in Table 3.3, 3.4 and 3.5.

If we examine correlation for answers of evaluators, we observe that evaluators answered more similarly for newsworthiness. The second similar one is responds for importance label and the least correlated answers are for correctness. As seen in the tables, correlations for newsworthiness and importance labels are high enough for ground truth, however, detection of correctness is a harder problem as observed in the lower values in the correlation.

Table 3.3: User answer correlation for newsworthiness

Correlation	**User 1**	**User 2**	**User 3**	**User 4**
User 1	100.0	78.36	83.64	84.28
User 2	78.36	100.0	84.32	81.88
User 3	83.64	84.32	100.0	86.80
User 4	84.28	81.88	86.80	100.0

Table 3.4: User answer correlation for importance

Correlation	**User 1**	**User 2**	**User 3**	**User 4**
User 1	100.0	65.16	64.28	78.64
User 2	65.16	100.0	78.00	71.32
User 3	64.28	78.00	100.0	69.40
User 4	78.64	71.32	69.40	100.0

Table 3.5: User answer correlation for correctness

Correlation	**User 1**	**User 2**	**User 3**	**User 4**
User 1	100.0	75.32	71.84	40.44
User 2	75.32	100.0	80.28	49.00
User 3	71.84	80.28	100.0	48.28
User 4	40.44	49.00	48.28	100.0

In order to construct the ground truth, we have used four different methods. Therefore, we have four different ground truths: GT_{avg}, GT_{4YES}, GT_{3YES}, GT_{1YES}.

- GT_{avg}**:** In this method, each YES answer is assigned two points, NEUTRAL answer is assigned one point and NO answer is assigned zero point. Four evaluations' scores are accumulated and total score is calculated. A tweet may have score between zero and eight points. If all of four evaluators answered YES for

a criterion, then total score for this criterion will be eight. If one of the evaluators answered YES and one of the evaluators answered NEUTRAL and two of evaluators answered NO, then total score for this criterion will be three. We set the value threshold as four. If total score is more than four, we classified these tweet as YES, otherwise as NO for the given criterion.

- GT_{4YES}**:** If a tweet is answered YES by all four users for the given criterion, then we classified this tweet as YES for this criterion. Otherwise, we classified it as NO.

- GT_{3YES}**:** If a tweet is answered YES by at least three users for the given criterion, then we classified this tweet as YES for this criterion. Otherwise, we classified it as NO.

- GT_{1YES}**:** This is the most relaxed form of ground truth. If a tweet is answered YES by at least one user for the given criterion, then we classified this tweet as YES for this criterion. Otherwise, we classified it as NO.

Note that we do not use GT_{2YES} since its mechanism is similar to GT_{avg}.

You can see the percentage of 2500 tweets that assigned YES under GT_{1YES}, GT_{3YES} and GT_{4YES} in the Table 3.6. As observed in the table, most of the tweets are evaluated as not newsworthy. Approximately %60 of tweets is answered as not newsworthy by all four users. The number of important tweets is more than number of newsworthy tweets. Approximately %66 of tweets is answered as important by at least one evaluator. Further, most of the tweets are evaluated as correct. Almost all tweets are answered as correct by at least one evaluator.

Table 3.6: Percentage of GT_{1YES}, GT_{3YES} and GT_{4YES} in all answers

Percentage	**Newsworthiness**	**Importance**	**Correctness**
GT_{1YES}	40.04	66.72	99.52
GT_{3YES}	19.00	28.48	76.64
GT_{4YES}	11.28	15.44	57.52

As also described in Section 3.1, we use the feature based learning approach as the first step. We train our data set with many learning schemes and try to found a model to predict the value of tweet in terms of three dimensions which are newsworthiness, importance and correctness. Firstly, we describe all features we used in our prediction system in Section 3.3.1. Then we will present classification details in Section 3.3.2.

Table 3.7: Overlap results of user answers and ground trurth resuls for newsworthiness label

Newsworthiness	GT_{avg}	GT_{4YES}	GT_{3YES}	GT_{1YES}
User 1	0.8920	0.8820	0.8916	0.7732
User 2	0.8764	0.8124	0.8680	0.9000
User 3	0.9320	0.8828	0.9352	0.8204
User 4	0.9168	0.8568	0.9116	0.8548
Average	0.9043	0.8585	0.9016	0.8371

Table 3.8: Overlap results of user answers and ground trurth resuls for importance label

Importance	GT_{avg}	GT_{4YES}	GT_{3YES}	GT_{1YES}
User 1	0.8364	0.7964	0.8364	0.6740
User 2	0.7808	0.6800	0.7776	0.8064
User 3	0.7856	0.6928	0.7832	0.7928
User 4	0.9044	0.8676	0.9052	0.6196
Average	0.8268	0.7592	0.8256	0.7232

Table 3.7, 3.8 and 3.9 display overlap results of user values and ground truth values. As seen in tables, GT_{avg} results match with user answers with high rate for each label. Some of users have better agreement rate with other ground truths but in average, user answers have the highest agreement rate with GT_{avg}. For newsworthiness 90.43 % of user answers match with GT_{avg}, for importance 82.68 % of user answers match with GT_{avg} and for correctness 85.17 % of user answers match with GT_{avg}. We also calculated Cohen's Kappa measurment which is presented in Table 3.10, 3.11 and 3.12. We calculated Kappa values with Equation 3.1. Since Kappa coefficient takes

Table 3.9: Overlap results of user answers and ground trurth resuls for correctness label

Correctness	GT_{avg}	GT_{4YES}	GT_{3YES}	GT_{1YES}
User 1	0.8276	0.7344	0.8208	0.7724
User 2	0.9164	0.7344	0.9196	0.7668
User 3	0.8732	0.8140	0.8684	0.7376
User 4	0.7896	0.5864	0.7684	0.9784
Average	0.8517	0.7353	0.8443	0.8138

into account random occurrences, it may give us a more robust results. We obtained the highest Kappa agreement with GT_{avg} for newsworthiness and importance labels and GT_{3YES} for correctness label.

$$[ht]K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3.1}$$

Table 3.10: Kappa results of user answers and ground trurth resuls for newsworthiness label

Newsworthiness	GT_{avg}	GT_{4YES}	GT_{3YES}	GT_{1YES}
User 1	0.6670	0.5672	0.6591	0.4863
User 2	0.6752	0.4569	0.6492	0.7827
User 3	0.7981	0.5941	0.8041	0.5970
User 4	0.7649	0.5394	0.7464	0.6783
Average	0.7257	0.5352	0.7135	0.6378

3.3 Feature Based Approach

In this phase, we train our data with several learning schemes and try to build a model to predict credibility. Firstly, we describe all features we used in this thesis. Then we will explain detail of our work for classification.

Table 3.11: Kappa results of user answers and ground trurth resuls for importance label

Importance	GT_{avg}	GT_{4YES}	GT_{3YES}	GT_{1YES}
User 1	0.6259	0.4872	0.6247	0.4071
User 2	0.5519	0.3363	0.5451	0.6194
User 3	0.5566	0.3506	0.5513	0.5961
User 4	0.7670	0.6245	0.7678	0.3341
Average	0.6194	0.4350	0.6159	0.4833

Table 3.12: Kappa results of user answers and ground trurth resuls for correctness label

Correctness	GT_{avg}	GT_{4YES}	GT_{3YES}	GT_{1YES}
User 1	0.4609	0.4153	0.4695	-0.1465
User 2	0.7561	0.5792	0.7757	0.0158
User 3	0.6416	0.5981	0.6428	-0.0275
User 4	0.0204	0.0326	0.0481	0.0537
Average	0.5182	0.4133	0.5239	-0.0436

3.3.1 Features

The features are grouped as tweet, user and topic features. In grouping and selecting the features, we followed the trend in the literature [6]. These features are used for building a model to predict label of a tweet.

We have total 41 features which contains 5 user features, 17 tweet features and 19 topic features.

User features are collected by using Twitter API. Then we calculated value of tweet features by simple string parsing operations. For sentiment score calculation, SentiStrength libraries [27] are used to calculate negative and positive sentiment score. At the end, we calculated the average scores for each topic. There are 100 tweets in each topic and average of features for 100 tweets computed for each topic.

Table 3.13: Descriptions of user features

Feature ID	**Feature Name**	**Description**
F_{U1}	Registration age	How long has it been since user registered to Twitter
F_{U2}	Number of total post	How many tweets are posted by user in total
F_{U3}	Number of friends	How many users are followed by this user
F_{U4}	Has description	User has description in his/her profile or not
F_{U5}	Has URL	User has URL in his/her profile or not.
F_{U6}	Is verified user	Is user a verified user by Twitter.

Table 3.14: Descriptions of tweet features

Feature ID	**Feature Name**	**Description**
F_{Tw1}	Tweet length	How many characters exist in the tweet? At most it can be 140 characters
F_{Tw2}	Number of words	How many words exist in the tweet
F_{Tw3}	Question mark	Presence of question mark in the tweet
F_{Tw4}	Exclamation mark	Presence of exclamation mark in the tweet
F_{Tw5}	Multiple mark	Presence of multiple question mark or exclamation mark in the tweet
F_{Tw6}	Contains smile	Presence of one of the smile icon in the tweet
F_{Tw7}	Contains frown	Presence of one of the frown icon in the tweet
F_{Tw8}	First pronoun	Presence of words for first pronoun
F_{Tw9}	Second pronoun	Presence of words for second pronoun
F_{Tw10}	Uppercase letters	Fraction of uppercase letters among all letters
F_{Tw11}	Contains URL	Presence of a URL in the tweet
F_{Tw12}	Mention character	Presence of @ mention tag in the tweet
F_{Tw13}	Hashtag character	Presence of # hashtag in the tweet
F_{Tw14}	Retweet	Is tweet a retweet or not
F_{Tw15}	Positive sentiment	How many positive words exist in the tweet?
F_{Tw16}	Negative sentiment	How many negative words exist in the tweet?
F_{Tw17}	Total sentiment	Sum of positive and negative sentiment score
F_{Tw18}	Weekday	Is tweet posted on Monday, Tuesday, Wednesday, Thursday or Friday

Table 3.15: Descriptions of topic features

Feature ID	**Feature Name**	**Description**
F_{Top1}	Average follower count	Topic average for number of followers per user
F_{Top2}	Average friend count	Topic average for number of friend per user
F_{Top3}	Average total post count	Topic average for number of post per user
F_{Top4}	Average registration age	Topic average for registration age per user
F_{Top5}	Fraction of smile	Topic average for presence of smile icon per tweet
F_{Top6}	Fraction of frown	Topic average for presence of frown icon per tweet
F_{Top7}	Fraction of hashtag	Topic average for presence of hashtag per tweet
F_{Top8}	Fraction of mention	Topic average for presence of mention per tweet
F_{Top9}	Fraction of exclamation mark	Topic average for presence of exclamation mark per tweet
F_{Top10}	Fraction of question mark	Topic average for presence of question mark per tweet
F_{Top11}	Fraction of multiple mark	Topic average for presence of multiple mark per tweet
F_{Top12}	Fraction of retweets	Average of retweets among all tweets in the topic
F_{Top13}	Fraction of description	Topic average for having description per user
F_{Top14}	Fraction of URL in tweet	Topic average for presence of URL per tweet
F_{Top15}	Fraction of URL in profile	Topic average for having URL in profile per user
F_{Top16}	Fraction of first pronoun	Topic average for presence of first pronoun per tweet
F_{Top17}	Fraction of second pronoun	Topic average for presence of second pronoun per tweet
F_{Top18}	Fraction of uppercase letters	Topic average for uppercase letters in tweet
F_{Top19}	Average length	Topic average for length of tweet

Table 3.13, 3.14 and 3.15 represent explanation of user, tweet and topic features. We eleminated one of the user feature which is "verified user" and one of the tweet features which is "weekday" since they do not have enough different value in our data set.

The details of the features in terms of their types, minimum values, maximum values and average values are listed in Table 3.16, 3.18 and 3.18.

Table 3.16: Details of user features

Feature Name	**Type**	**Min. Value**	**Max. Value**	**Mean**
Registration age	Number	0	2214	570.2
Number of total post	Number	1.0	262104	5852.2
Number of friends	Number	0	93794	553.3
Has description	Bool	-	-	-
Has URL	Bool	-	-	-

Table 3.17: Details of tweet features

Feature Name	**Type**	**Min. Value**	**Max. Value**	**Mean**
Tweet length	Number	9	160	100
Number of words	Number	1	41	13
Question mark	Bool	-	-	-
Exclamation mark	Bool	-	-	-
Multiple mark	Bool	-	-	-
Contains smile	Bool	-	-	-
Contains frown	Bool	-	-	-
First pronoun	Bool	-	-	-
Second pronoun	Bool	-	-	-
Uppercase letters	Number	0	0.87	0.09
Contains URL	Bool	-	-	-
Mention character	Bool	-	-	-
Hashtag character	Bool	-	-	-
Retweet	Bool	-	-	-
Positive sentiment	Number	1	5	1.5
Negative sentiment	Number	-5	-1	-1.3
Total sentiment Number	-4	4	0.1	

Table 3.18: Details of topic features

Feature Name	**Type**	**Min. Value**	**Max. Value**	**Mean**
Average follower count	Number	277.9	19947	3074.6
Average friend count	Number	268.7	1356.4	653.68
Average total post count	Number	2471.8	18501.7	7034.5
Average registration age	Number	166.16	1082.63	655.81
Fraction of smile	Number	0	0.3	0.1
Fraction of frown	Number	0	0.2	0.01
Fraction of hashtag	Number	0	1	0.376
Fraction of mention	Number	0.2	1	0.724
Fraction of excl. mark	Number	0.02	1.22	0.172
Fraction of quest. mark	Number	0	0.2	0.1
Fraction of multiple mark	Number	0	0.8	0.1
Fraction of retweets	Number	0	1	0.39
Fraction of description	Number	0.14	1	0.84
Fraction of URL in tweet	Number	0	0.6	0.3
Fraction of URL in profile	Number	0.04	0.81	0.35
Fraction of first pronoun	Number	0	0.41	0.06
Fraction of second pronoun	Number	0	0.2	0.05
Fraction of uppercases	Number	0.04	0.25	0.11
Average length	Number	73.78	320.9	119.5

Figure 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 gives histogram details of some important features for newsworthiness label. Red color represents newsworthy tweets and blue color represents not newsworthy tweets. Histogram details for rest of the features are presented in Appendix. Appendix A contains figure for newsworthiness label, Appendix B contains figure for importance label and Appendix C contains figure for correctness label.

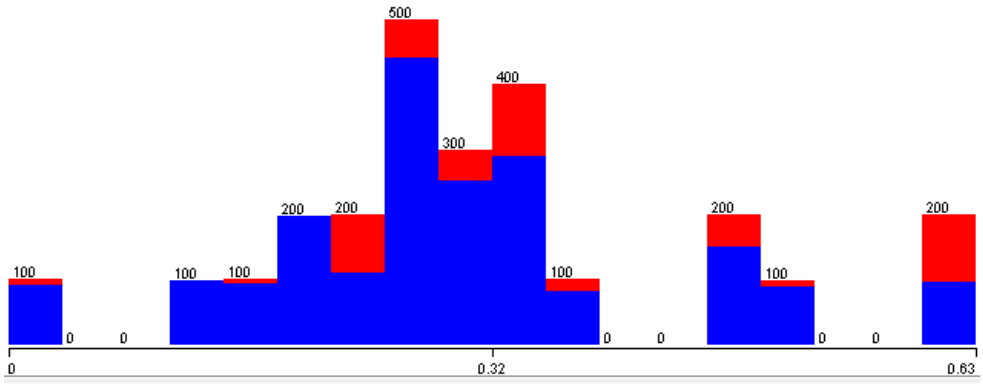

Figure 3.3: Histogram Detail of Feature - Fraction of URL in profile

Fraction of URL in profile is a topic feature. As presented in Figure 3.3, if more authors in topic have URL in their profile, newsworthy tweets are more likely to be included in this topic.

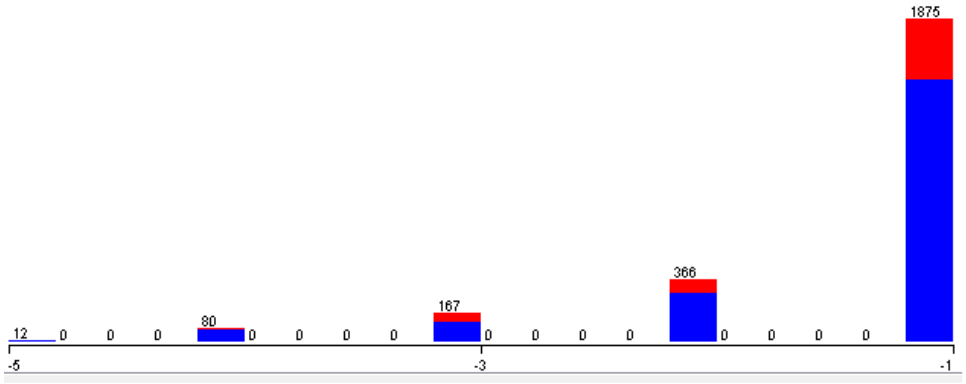

Figure 3.4: Histogram Detail of Feature - Negative Sentiment Score

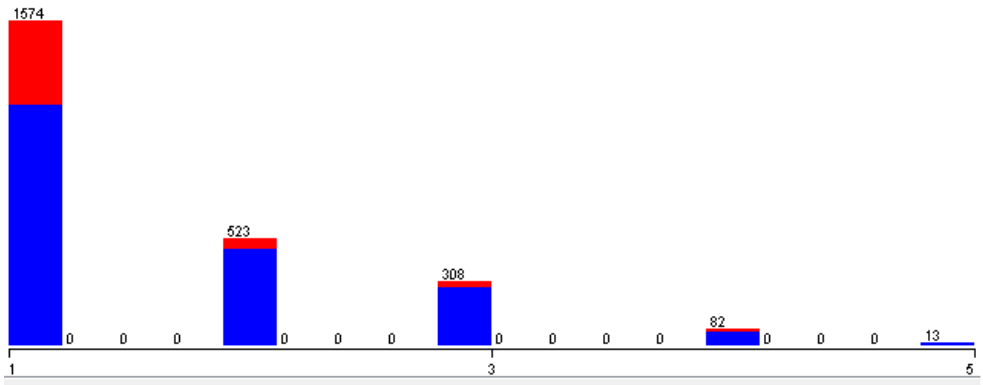

Figure 3.5: Histogram Detail of Feature - Positive Sentiment Score

Negative Sentiment Score is a tweet feature and represents number of negative word in the post. Figure 3.4 displays that tweets, which have less negative words, are more likely to be newsworthy. Similar to negative sentiment, tweets containing less positive words, are also more likely to be newsworthy as seen in Figure 3.5. We can conclude that tweets having more emotional words are less likely to be newsworthy.

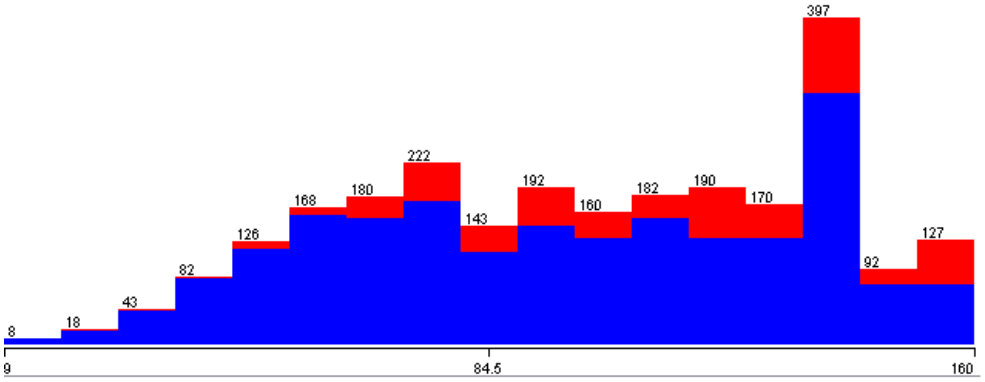

Figure 3.6: Histogram Detail of Feature - Tweet Length

Tweet Length is a tweet feature. As presented in Figure 3.6, longer tweets are more likely to be newsworthy.

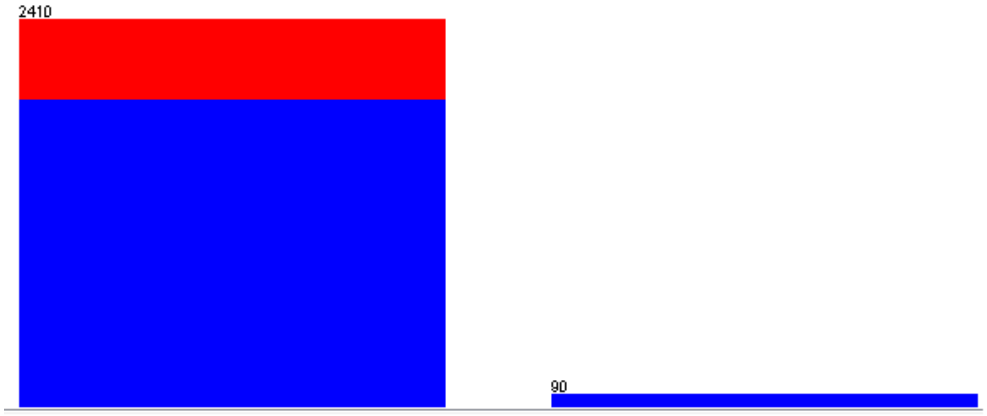

Figure 3.7: Histogram Detail of Feature - Second Pronoun

If a tweet contains second pronoun words, it is less likely to be newsworthy. As seen in Figure 3.7, tweets containing second pronoun are not newsworthy tweets.

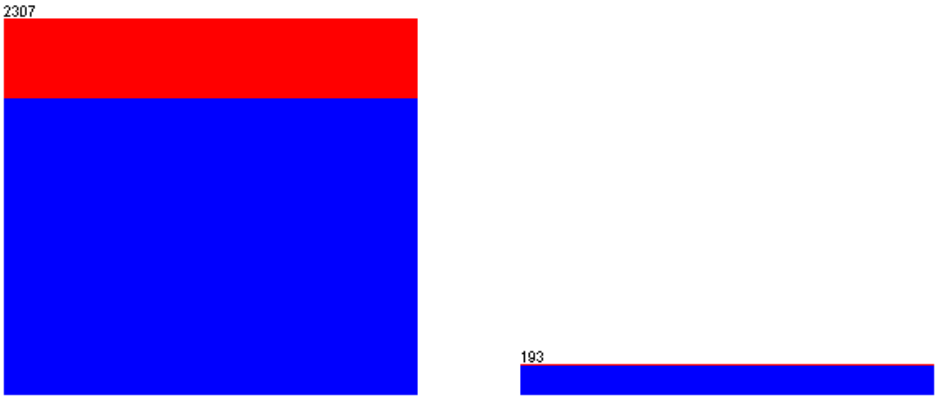

Figure 3.8: Histogram Detail of Feature - Contains smile

Figure 3.8 presents us that tweets having smile icons are less likely to be newsworthy.

3.3.2 Classification

Some of features do not have sufficient number of instances with different values. Therefore, we eliminated these features since they have no impact on determining class. These features are weekday, verified user and average verified user in topic. After we have collected all attributes for tweet, user and topic, we calculated an initial score by using different decision algorithms. We used KNIME[18] and Weka[36] data analysis tools to visualize different classifiers.

3.3.2.1 Classification with KNIME Tool

We tried decision trees, naive bayes and SVM by using KNIME[18] application. Result of these learning schemes will be explained in Chapter 4. Figure 3.9 displays learning schemes in KNIME tool.

3.3.2.2 Classification with Weka Tool

We can try much more algorithms with Weka [36] application. We used 10 folds cross validation in each learning schemes. GT_{avg} is used in all of our experiment

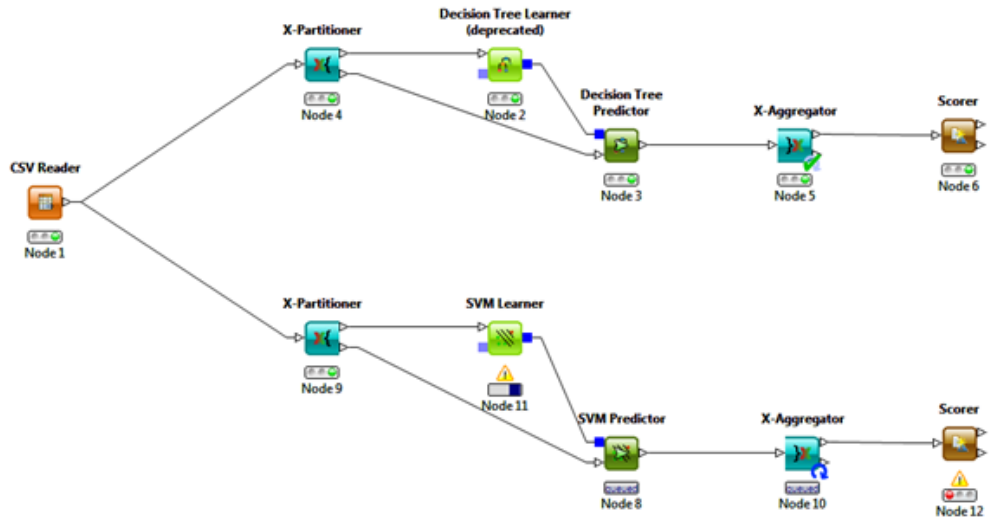

Figure 3.9: Learning Schemes for Knime Tool

with Weka tool and we applied same process for newsworthiness, importance and correctness labels.

When predicting labels, we used two different methods as follow.

- We used all 43 features together to generate the model.
- We grouped the features as user features, tweet feature and topic features. Then we used each type of features separately to generate a model. Therefore, tweet features are used to predict labels of tweets, user features are used to predict label of users and topic features are used to predict labels of topics.

We used following learning schemes with Weka tool.

- Random Forest Tree
- J48 Tree
- ADTree
- Random Tree
- BFTree
- Naive Bayes

- KStar
- AdaBoost

Prediction results with Weka tool are explained in Chapter 4.

3.4 Graph Based Approach

After feature based learning, we applied authority transfer between tweets, users and topics. At the end of the first phase, it is possible that a credible user may have low score. However, if this user has important tweets, then authority transfer enables that important tweets make their author more important. Similarly if an unreliable user has high score after the first phase, if this user has unimportant tweets, these tweets make author also less important. Furthermore, if we go through tweet - topic relationship, we see that if a topic has important tweets, tweets in this topic will also be more important. Similarly if a tweet is important, it makes its topic more important.

In our data set each tweet has a topic and a user. Figure 3.4 displays graph structure used in the authority transfer step of our solution.
Nodes are:

- Tweet
- User
- Topic

Edges are:

- Tweet to User Edge
- User to User Edge
- Tweet to Topic Edge
- Topic to Tweet Edge

Figure 3.10: Graph Structure of Our Study

Score of tweet node has influence on user and topic nodes. User nodes firstly affects tweet nodes then it affects topic nodes indirectly by transferring score from tweet to topic. Topic nodes send their scores to tweets and they also send their scores indirectly to users in the second step. Hence, these transfers implement the following effects.

- A tweet is more important if its author is important.
- A tweet is more important if its topic is important.
- An author is important if he/she posted important tweets.
- A topic is important if it contains important tweet.

We have initial scores for each tweet, user and topic from the result of the feature based learning phase. The number of retweets of a tweet is added to the feature score of each tweet node and the number of followers is added to the feature score of each user node.

Table 3.19: Definitions of variables in equations

Name	**Definition**
$S_{\#fol}$	Number of followers for each user
$S_{\#rt}$	Number of retweet for each tweet
w_1	Weight for User to Tweet edge
w_2	Weight for Tweet to User edge
w_3	Weight for Topic to Tweet edge
w_4	Weight for Tweet to Topic edge

In our model, authority transfer is evaluated by using the Equations 3.5, 3.6 and 3.7.

Definition of variables in equations are given in Table 3.19.

$$S_{tweet0} = S_{feature} + S_{\#rt} \tag{3.2}$$

$$S_{user0} = S_{feature} + S_{\#fol} \tag{3.3}$$

$$S_{topic0} = S_{feature} \tag{3.4}$$

$$S_{tweet} = S_{tweet0} + w_1 * S_{user0} + w_3 * S_{topic0} \tag{3.5}$$

$$S_{user} = S_{user0} + ((w_3 * S_{topic0}) + S_{tweet0}) * w_2 \tag{3.6}$$

$$S_{topic} = S_{topic0} + ((w_1 * S_{user0}) + S_{tweet0}) * w_2 \tag{3.7}$$

The final score of a tweet is the sum of initial score of this tweet, score coming from user and score coming from topic. When we calculate the final score of a user, firstly topic score is transferred to tweet's initial score. Then tweet score is added to the initial score of this user. For the final score of topic score, firstly user score is transferred to tweet and then tweet score is added to initial score of this topic. At the end, we have final scores for each tweet, user and topic.

We can conclude that if the final score of a tweet is more than a predefined threshold, then tweet is newsworthy or important or correct.

After trying several threshold values, we get maximum accuracy rate when threshold is 10. We calculated score of tweets for each of label and we conclude that

- If newsworthiness score of a tweet is less than 10, the tweet is not newsworthy otherwise it is newsworthy.

- If importance score of a tweet is less than 10, the tweet is not important otherwise it is important.
- If correctness score of a tweet is less than 10, the tweet is incorrect otherwise it is correct.
- Threshold: 10

Scores are transmitted among the nodes according to values of weights. We experimented with several sets of weights and get the best result with following coefficient set.

- w_1: 0.1
- w_2: 10
- w_3: 1
- w_4: 10
- Threshold: 10

CHAPTER 4

EXPERIMENTAL EVALUATION

In this chapter, we will describe results of our experiments. Section 4.1 gives result of the feature based training which is our first phase process. Section 4.2 explains the result of prediction after authority transfer.

4.1 Result of Feature Based Learning

We used KNIME[18] and Weka[36] data analysis tools to visualize different learning schemes.

4.1.1 Experiment Results by Using KNIME Tool

We only calculated score for newsworthiness label by using KNIME tool since we obtained better results with Weka Tool. We tried decision trees, naive bayes and SVM by using KNIME[18] tool. The best result is given by using decision tree.

We obtained the following results.

Newsworthiness: The best accuracy rate is 87.44 % for newsworthy class with decision tree learner.
Importance: The best accuracy rate is 83.68 % for importance class with decision tree learner.
Correctness: The best accuracy rate is 79.32 % for correctness class with decision tree learner.

4.1.2 Experiment Results by Using Weka Tool

We can try much more algorithms with Weka application and results from different learning schemes are listed in Table 4.1 4.2 and 4.3. We used 10 folds cross validation in each learning schemes. GT_{avg} is used in all of our experiment with Weka tool.

For newsworthy label, we get the best accuracy result from random forest decision tree learner.
89.64 % of 2500 tweets are classified correctly by using all features.
82.68 % of 2500 tweets are classified correctly by using only user related features.
87.56 % of 2500 tweets are classified correctly by using only tweet related features.
85.84 % of 2500 tweets are classified correctly by using only topic related features.

Table 4.1: Newsworthiness label accuracy rate result of classification algorithms

Accuracy Rate	**All Feat.**	**User Feat.**	**Tweet Feat.**	**Topic Feat.**
Random Forest Tree	89.64	82.68	87.56	85.84
J48 Tree	89.44	80.24	86.28	85.84
ADTree	87.80	79.72	83.16	85.84
Random Tree	85.24	76.32	85.16	85.84
BFTree	89.04	79.84	86.16	85.84
Naive Bayes	75.96	79.44	82.04	65.36
KStar	84.92	80.68	86.96	85.84
AdaBoost	85.28	79.92	81.52	83.92

For importance label, we get the best accuracy result from random forest decision tree learner.
83.52 % of 2500 tweets are classified correctly by using all features.
75.84 % of 2500 tweets are classified correctly by using only user related features.
81.20 % of 2500 tweets are classified correctly by using only tweet related features.
81.32 % of 2500 tweets are classified correctly by using only topic related features.

Table 4.2: Importance label accuracy rate result of classification algorithms

Accuracy Rate	**All Feat.**	**User Feat.**	**Tweet Feat.**	**Topic Feat.**
Random Forest Tree	83.52	75.84	81.20	81.32
J48 Tree	82.40	73.44	79.48	81.08
ADTree	83.36	72.60	73.64	81.32
Random Tree	80.52	70.08	77.92	81.32
BFTree	82.88	72.48	78.44	81.32
Naive Bayes	81.12	71.20	73.12	80.64
KStar	81.56	73.40	81.04	81.32
AdaBoost	83.08	70.92	74.96	80.16

For correctness label, we get the best accuracy result from random forest decision tree learner for all features, tweet features and topic features. However, we get the best accuracy result from J48 decision tree for user features.
82.72 % of 2500 tweets are classified correctly by using all features.
79.88 % of 2500 tweets are classified correctly by using only user related features.
81.40 % of 2500 tweets are classified correctly by using only tweet related features.
81.88 % of 2500 tweets are classified correctly by using only topic related features.

Table 4.3: Correctness label accuracy rate result of classification algorithms

Accuracy Rate	**All Feat.**	**User Feat.**	**Tweet Feat.**	**Topic Feat.**
Random Forest Tree	82.72	79.36	81.40	80.72
J48 Tree	80.48	79.88	80.84	81.88
ADTree	79.24	79.44	81.12	81.88
Random Tree	78.32	73.72	76.24	81.88
BFTree	81.40	79.60	80.96	81.88
Naive Bayes	66.76	26.44	77.68	77.60
KStar	78.82	75.72	79.56	81.88
AdaBoost	81.16	79.36	80.00	81.88

4.2 Result of Authority Transfer

We experiment authority transfer for three labels, which are newsworthiness, importance and correctness. In this chapter, we will describe experiment result for each label one by one.

While we evaluate the results, we regard a tweet as positive label with four different interpretations which are GT_{avg}, GT_{4YES}, GT_{3YES} and GT_{1YES}. Explanation of these interpretations is described in Section 3.2.

We tried two different ways to calculate initial score for nodes. The first way is giving all 41 features together to decision tree learner. Then each type of nodes is started with initial score of prediction with decision tree learner. The second way is giving features separately such as user related ones, tweet related ones and topic related ones. So after prediction results, each node has its own initial score from prediction results of its own type. For example, prediction results of user related features have only influence on user nodes.

4.2.1 Newsworthiness Label

Our first experiment is using GT_{avg} evaluation score and all features together. Prediction result from random forest decision tree was 89.64 %. After authority transfer success rate increased to 90.92 %.

We applied various filtering methods to data set. However, most of them did not give good result. Two of the filtering methods have limited positive effect on success rate. These are zero filtering and interquartile range filtering methods.

If we use zero filtering prediction result from random forest decision tree increases to 89.96 % and after authority transfer it reaches 90.96 %.

Another filtering method which we tried is interquartile range. It raised success rate of random forest tree to 89.88 % and after authority transfer success rate reached 91.12 %. Since filtering doesn't change success rate enough, we will not use it for other labels. You can see comparision of prediction accuracy rate in Figure 4.1

Figure 4.1: Comparision of filtering methods for newsworthiness accuracy rate

When we used GT_{4YES} as a positive feedback from evaluators, success rate becomes 89.16 % after authority transfer. It is 90.96 % for GT_{3YES} and 83.16 % for GT_{1YES}. We obtain the best accuracy rate with GT_{3YES} evaluation. Figure 4.2 represents accuracy rate in GT_{avg}, GT_{4YES}, GT_{3YES} and GT_{1YES}.

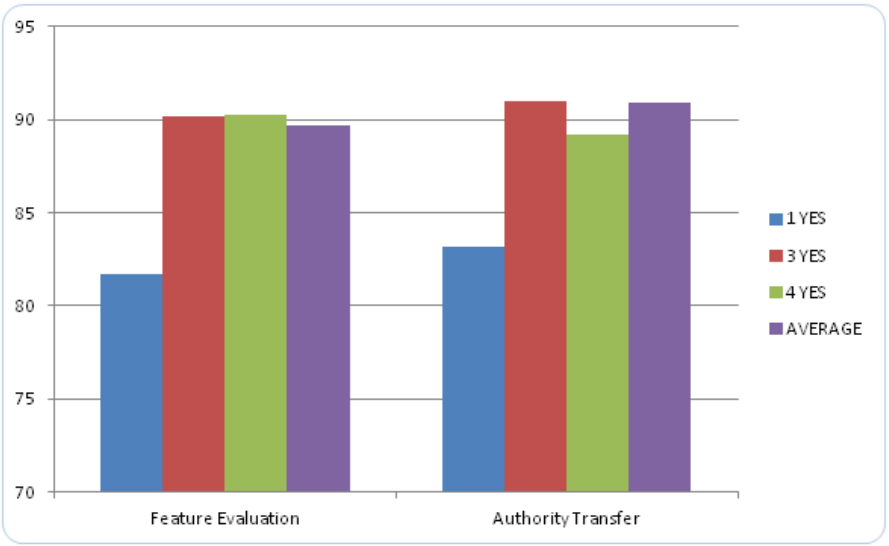

Figure 4.2: Newsworthiness label accuracy rate when training all features together

If we experiment GT_{avg} by using grouped features, we reach 91.8 % successful prediction rate which is slightly more than experiment result with all features together. If we regard tweets newsworthy with GT_{4YES} ground truth, success rate goes to 89.88 %. For GT_{3YES}, it goes to 84.64 % and for GT_{1YES}, it goes to 83.88 %. We get the best success rate when we evaluate a tweet is newsworthy for GT_{avg}.

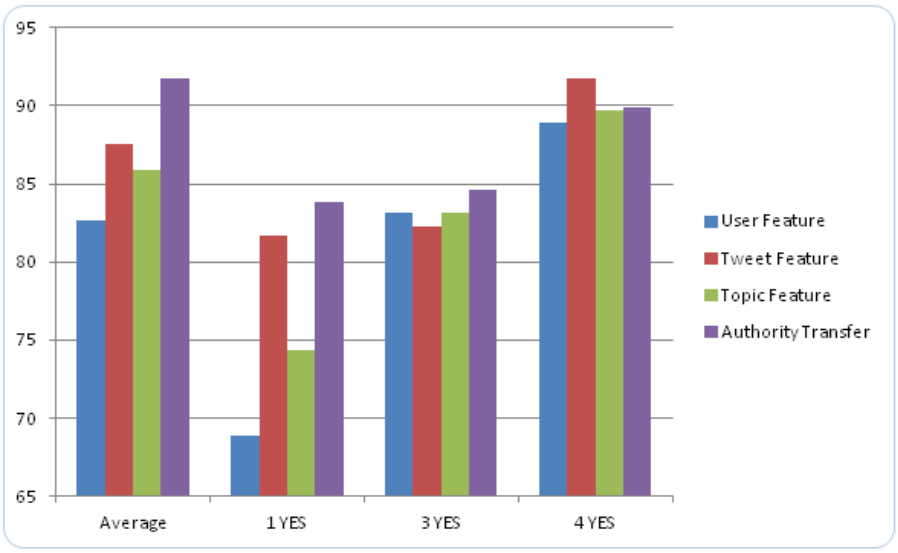

Figure 4.3: Newsworthiness label accuracy rate when training features separately

Figure 4.3 express that success rate increases slightly when we give prediction results separately to nodes as initial score.

We analyzed effect of number of topic on newsworthiness label. Figure 4.4 represents experiment results with different number of topics. You can see comparison of successful prediction accuracy of data sets having 10 topics, 15 topic and 25 topics in this figure. We selected topics randomly for each experiment. Each topic contains 100 tweets in all experiments and GT_{avg} is used in evaluation. Results show that having more topics in the data set increases accuracy results.

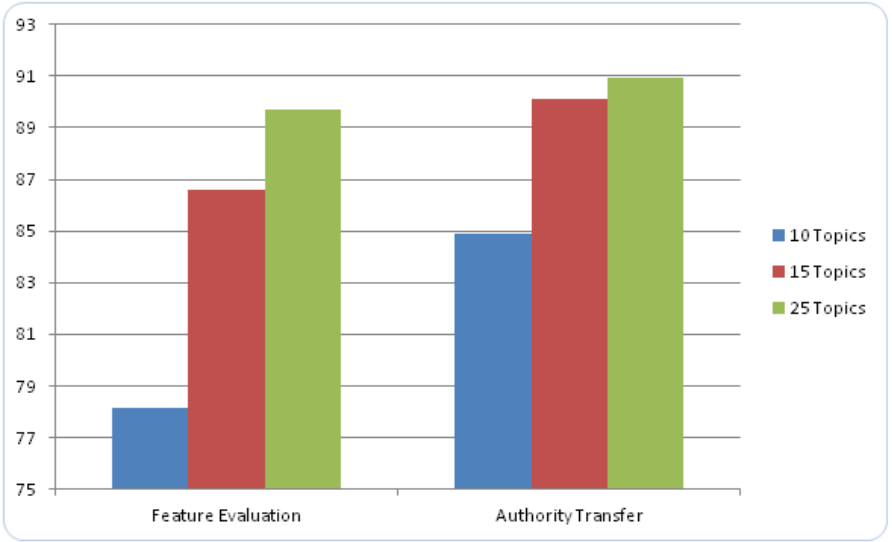

Figure 4.4: Comparision of accuracy rate for different number of topics

Figure 4.5 displays comparison of experiments with equal number of topics but different number of tweets in each topic. Each experiment has 25 topics but the first one has 50 tweets for topic, the second one has 80 tweets for topic and the third one has 100 tweets for topic. We determined tweets with a random selection from our data set. Results of this experiment show us that if we increase number of tweets in each topic, we can predict newsworthiness more successfully with GT_{avg}. We also see that authority transfer is more effective if the size of data is small.

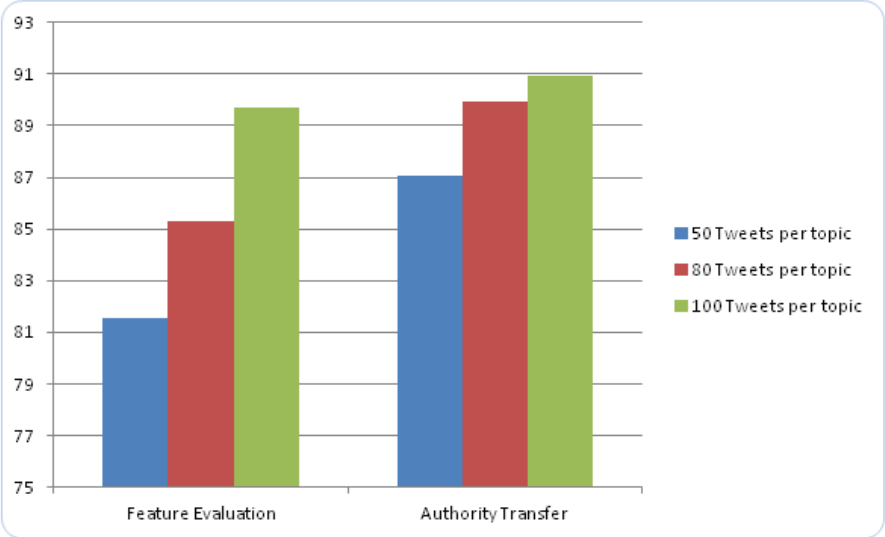

Figure 4.5: Comparision of accuracy rate for different number of tweets per topic

4.2.2 Importance Label

We used GT_{avg}, GT_{4YES} and GT_{3YES} for evaluation of importance label. We didn't use GT_{1YES} since its result is not in expected level.

Our initial experiment is using GT_{avg} with all features together. Success rate from random decision tree is 83.52 and after authority transfer it reaches 84.24 %. When evaluating with GT_{4YES}, success rate is 86.88 % and when evaluating with GT_{3YES}, success rate is 84.64 %. We get the best result with when we regard a tweet important if it is considered as important by four users as seen in Figure 4.6.

If we train features separately for user, tweet and topic then we get following results. For GT_{avg}, success rate is 82.76 %. We obtained 82.80 % success rate for GT_{3YES} evaluation and 84.84 % for GT_{4YES} evaluation as seen in Figure 4.7.

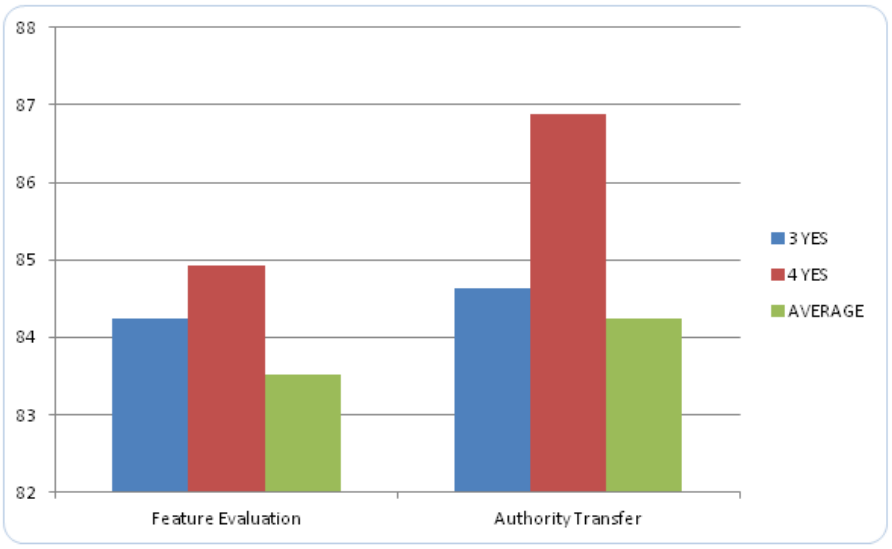

Figure 4.6: Importance label accuracy rate when training all features together

For importance label we get the best scores when we accept four important answers from evaluators.

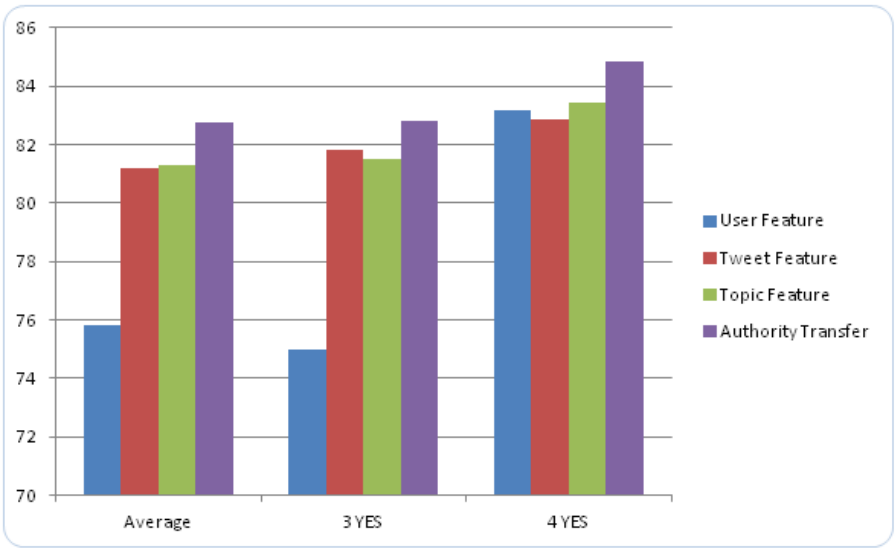

Figure 4.7: Importance label accuracy rate when training features separately

4.2.3 Correctness Label

When we are evaluating correctness class, we used GT_{avg}, GT_{4YES} and GT_{3YES} for correctness label.

Our initial experiment is using GT_{avg} with all features together. Success rate from

random decision tree is 82.72 % and after authority transfer it reaches 84.12 %. When evaluating with GT_{4YES}, success rate is 76.92 % and when evaluating with GT_{3YES}, success rate is 81.44 %. We get the best result when we regard a tweet correct according to GT_{avg} as seen in Figure 4.8.

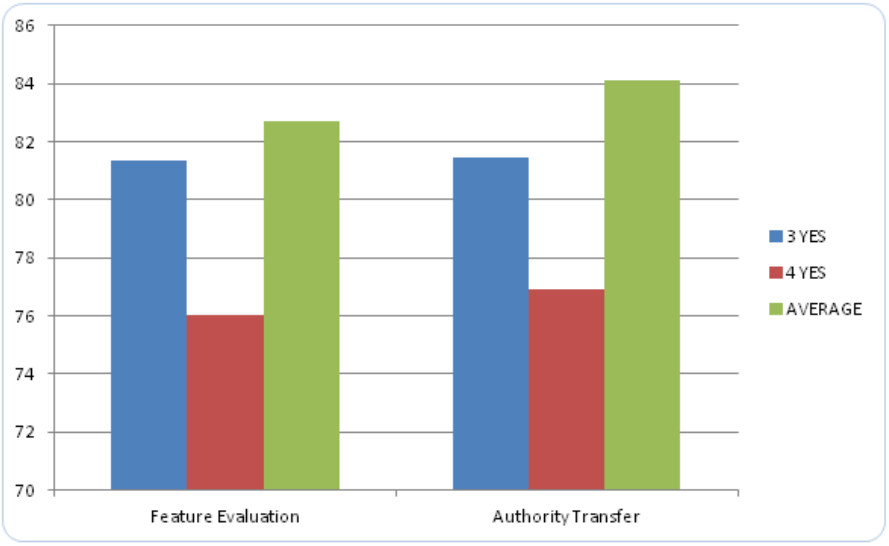

Figure 4.8: Correctness label accuracy rate when training all features together

If we use features separately for user, tweet and topic then we get following results. For GT_{avg}, success rate is 83.20 %. We obtained 79.20 % success rate for GT_{3YES} evaluation and 57.84 % for GT_{4YES} evaluation as seen in Figure 4.9.

For correctness label we get best scores when we use GT_{avg}.

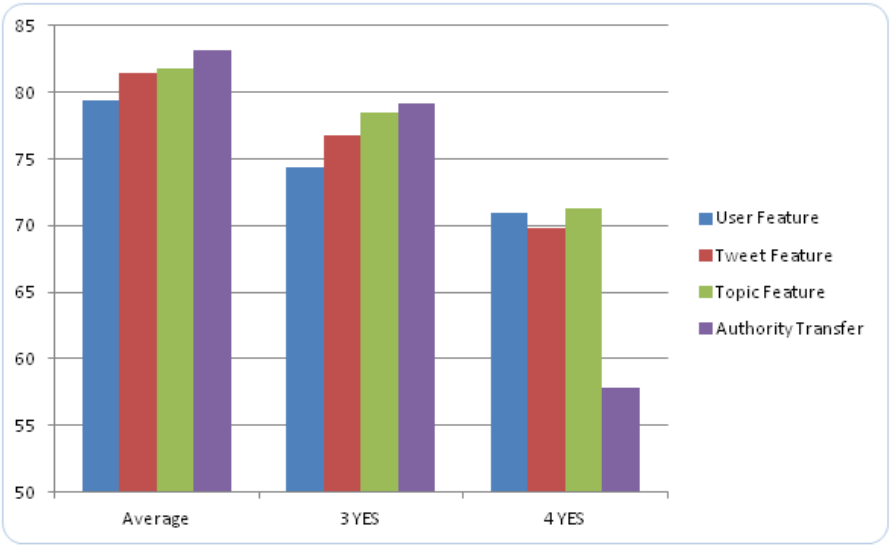

Figure 4.9: Correctness label accuracy rate when training features separately

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Widespread usage of microblogging services associates with credibility of data in microblogging services. In this thesis, we investigated Twitter which is one of the most popular Microblog. Many people use Twitter actively and their posts become valuable for some studies which are interested with huge information in Twitter. However, it brings to mind one problem which is credibility of data.

There exist some studies which are investigating credibility problem in Twitter. We can classify these studies into two groups which are feature based approaches and graph based approaches. We decided to use a hybrid model which utilizes from both solutions. Study of Castillo et al. [6] uses several features and trains them to predict credibility. On the other hand, Yamaguchi et al. [38] presents TURank algorithm to rank tweets in terms of their newsworthiness. A Page rank like algorithm is used in a graph of tweet and user nodes. We are impressed by studies of Castillo and Yamaguchi, and then presented a new approach which combines feature based approach and graph based approach.

We evaluated tweets in terms of three labels which are newsworthiness, importance and correctness. Credibility studies in literatures generally focus one of these three labels. We measure these labels together in this thesis. Initially, we asked our evaluators to get their value of these labels for each tweet. We used four different methods to consider value of evaluator feedbacks. These are GT_{avg}, GT_{4YES}, GT_{3YES}, GT_{1YES}. In the first phase, we trained our data set according to result of user feedbacks by

GT_{avg}, GT_{4YES}, GT_{3YES}, GT_{1YES}. We obtained the best prediction result with random forest decision tree. When we trained all 41 features together, we obtained 89.64 percent successful prediction rate for newsworthiness. 83.52 percent success rate is achieved for importance label with random forest decision tree and 84.80 percent success is obtained for correctness label. We also trained our data set by separating features for user related, tweet related and topic related. We applied some filtering methods before training data with random decision tree learner. However, we couldn't get a remarkable change with filtering methods.

In second phase, we used the prediction results in the first phase and transferred this score in our graph. We defined three types of nodes which are user, tweet and topic. Users transfer their score to their tweets. Topics transfer their scores to tweets and tweets transfer score to their authors and their topics. After authority transfer, we obtained successful prediction rate between 80 percent and 92 percent for a range of experiments. For newsworthiness label, we obtained best result with GT_{avg} evaluation method with 91.80 percent success rate. We achieved this score by training user, tweet and topic features separately. For importance label, best result is obtained with GT_{4YES} evaluation method. We predicted 86.88 percent of tweets correctly when we trained 41 features together. Lastly for correctness label, we obtained 84.12 percent success rate with GT_{avg} evaluation method when we used all feature together.

We see that if we apply authority transfer after feature based approach, prediction accuracy increases. We generally obtained better result with GT_{avg} evaluation method but for importance label we got best result with GT_{4YES} evaluation method. Training features separately for user, tweet and topic increases success rate for newsworthiness but it does not affect other labels positively.

5.2 Future Work

We can increase the size of our data set as a future work. It may provide us better prediction accuracy for credibility. If we increase size of our data set, we believe that authority transfer will be more effective because there will be more users and more tweets in the data set. So users will get score from their friends and tweets will get

their score from their retweets.

Behavioral features of users in Twitter can also help us to measure credibility. There exist some studies in literature which utilize from behavioral attributes such as frequency of updating profile, interaction with other users, retweeting frequency and retweeting timing [1, 19]. Using behavioral features may lead us to classify credibility of user more successfully.

Grammer rules and spelling of words may be good features for successful prediction. They can be included to feature set as a future work.

We can calculate better weights for edges in authority transfer. It may help us to increase success of prediction.

We applied our approach in Turkish tweets. However, the proposed technique is language independent. Hence, in the future, it is possible to use the same mechanism to tweets from other language. The only language dependent points in our approach is sentiment score calculation and personal pronouns processing in tweet features. We calculated sentiment score for Turkish language, but it is not a big effort to adapt it for another language for future work. Similarly we also need to adapt our system to find personal pronouns for other languages.

Our data is evaluated by 4 users. If we ask more users to label our tweets, it may lead us better results.

REFERENCES

[1] M.-A. Abbasi and H. Liu. Measuring user credibility in social media. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 441–448, Berlin, Heidelberg, 2013. Springer-Verlag.

[2] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, pages 1–12, Berlin, Heidelberg, 2011. Springer-Verlag.

[3] O. Alonso, C. C. Marshall, and M. Najork. Are some tweets more interesting than others? #hardquestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR '13, pages 2:1–2:10, New York, NY, USA, 2013. ACM.

[4] M. Armentano, D. Godoy, and A. Amandi. Recommending information sources to information seekers in twitter. In *International Workshop on Social Web Mining*, UMAP'11, 2011.

[5] M. Armentano, D. Godoy, and A. Amandi. Topology-based recommendation of users in micro-blogging communities. volume 27, pages 624–634. Springer US, 2012.

[6] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[7] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[8] Facebook. https://www.facebook.com. Last access: 5 Jan 2014.

[9] W. Feng and J. Wang. Retweet or not?: Personalized tweet re-ranking. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 577–586, New York, NY, USA, 2013. ACM.

[10] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8, New York, NY, USA, 2012. ACM.

[11] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter, 2012.

[12] H. Huang, A. Zubiaga, H. Ji, H. Deng, D. Wang, H. K. Le, T. F. Abdelzaher, J. Han, A. Leung, J. Hancock, and C. R. Voss. Tweet ranking based on heterogeneous networks. In *COLING*, pages 1239–1256, 2012.

[13] L. Huang and X. Yeming. Evaluation of microblog users' influence based on pagerank and users behavior analysis. In *Advances in Internet of Things*, pages 34–40, 2013.

[14] Instagram. http://instagram.com. Last access: 5 Jan 2014.

[15] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 657–664, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[16] B. Kang, J. O'Donovan, and T. Höllerer. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 179–188, New York, NY, USA, 2012. ACM.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999.

[18] Knime. http://www.knime.org/. Last access: 5 Jan 2014.

[19] S. Kong and L. Feng. A tweet-centric approach for topic-specific author ranking in micro-blog. In *Proceedings of the 7th International Conference on Ad-*

vanced Data Mining and Applications - Volume Part I, ADMA'11, pages 138–151, Berlin, Heidelberg, 2011. Springer-Verlag.

[20] M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[21] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[22] J. O'Donovan, B. Kang, G. Meyer, T. Höllerer, and S. Adalii. Credibility in context: An analysis of feature distributions in twitter. In *SocialCom/PASSAT*, pages 293–301. IEEE, 2012.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[24] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 45–54, New York, NY, USA, 2011. ACM.

[25] S. Ravikumar, R. Balakrishnan, and S. Kambhampati. Ranking tweets considering trust and relevance. In *Proceedings of the Ninth International Workshop on Information Integration on the Web*, IIWeb '12, pages 4:1–4:4, New York, NY, USA, 2012. ACM.

[26] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[27] Sentistrength library source. `http://sentistrength.wlv.ac.uk`. Last access: 5 Jan 2014.

[28] Y. Suzuki and A. Nadamoto. Credibility assessment using wikipedia for messages on social network services. In *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*, pages 887–894, 2011.

[29] Tumblr. `https://www.tumblr.com`. Last access: 5 Jan 2014.

[30] Twitter. `https://twitter.com`. Last access: 5 Jan 2014.

[31] Twitter API. `https://dev.twitter.com/docs`. Last access: 5 Jan 2014.

[32] I. Uysal and W. B. Croft. User oriented tweet ranking: A filtering approach to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2261–2264, New York, NY, USA, 2011. ACM.

[33] Vine. `https://vine.co`. Last access: 5 Jan 2014.

[34] W. Weerkamp and M. Rijke. Credibility-inspired ranking for blog post retrieval. volume 15, pages 243–277. Springer Netherlands, 2012.

[35] Weibo. `http://www.weibo.com`. Last access: 5 Jan 2014.

[36] Weka. `http://www.cs.waikato.ac.nz/ml/weka/`. Last access: 5 Jan 2014.

[37] X. Xia, X. Yang, C. Wu, S. Li, and L. Bao. Information credibility on twitter in emergency situation. In *Proceedings of the 2012 Pacific Asia Conference on Intelligence and Security Informatics*, PAISI'12, pages 45–59, Berlin, Heidelberg, 2012. Springer-Verlag.

[38] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In *Proceedings of the 11th International Conference on Web Information Systems Engineering*, WISE'10, pages 240–253, Berlin, Heidelberg, 2010. Springer-Verlag.

[39] Yammer. `https://www.yammer.com`. Last access: 5 Jan 2014.

[40] J. Yang, S. Counts, M. R. Morris, and A. Hoff. Microblog credibility perceptions: Comparing the usa and china. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 575–586, New York, NY, USA, 2013. ACM.

[41] M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim. Finding interesting posts in twitter based on retweet graph analysis. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1073–1074, New York, NY, USA, 2012. ACM.

[42] X. Zhang, S. Zhu, and W. Liang. Detecting spam and promoting campaigns in the twitter social network. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1194–1199, 2012.

APPENDIX A

HISTOGRAM OF NEWSWORTHINESS LABEL

You can see histogram of newsworthiness label in Figure A.1, A.2 and A.3. Red color represents true newsworhty tweets and blue color represents not newsworthy tweets.

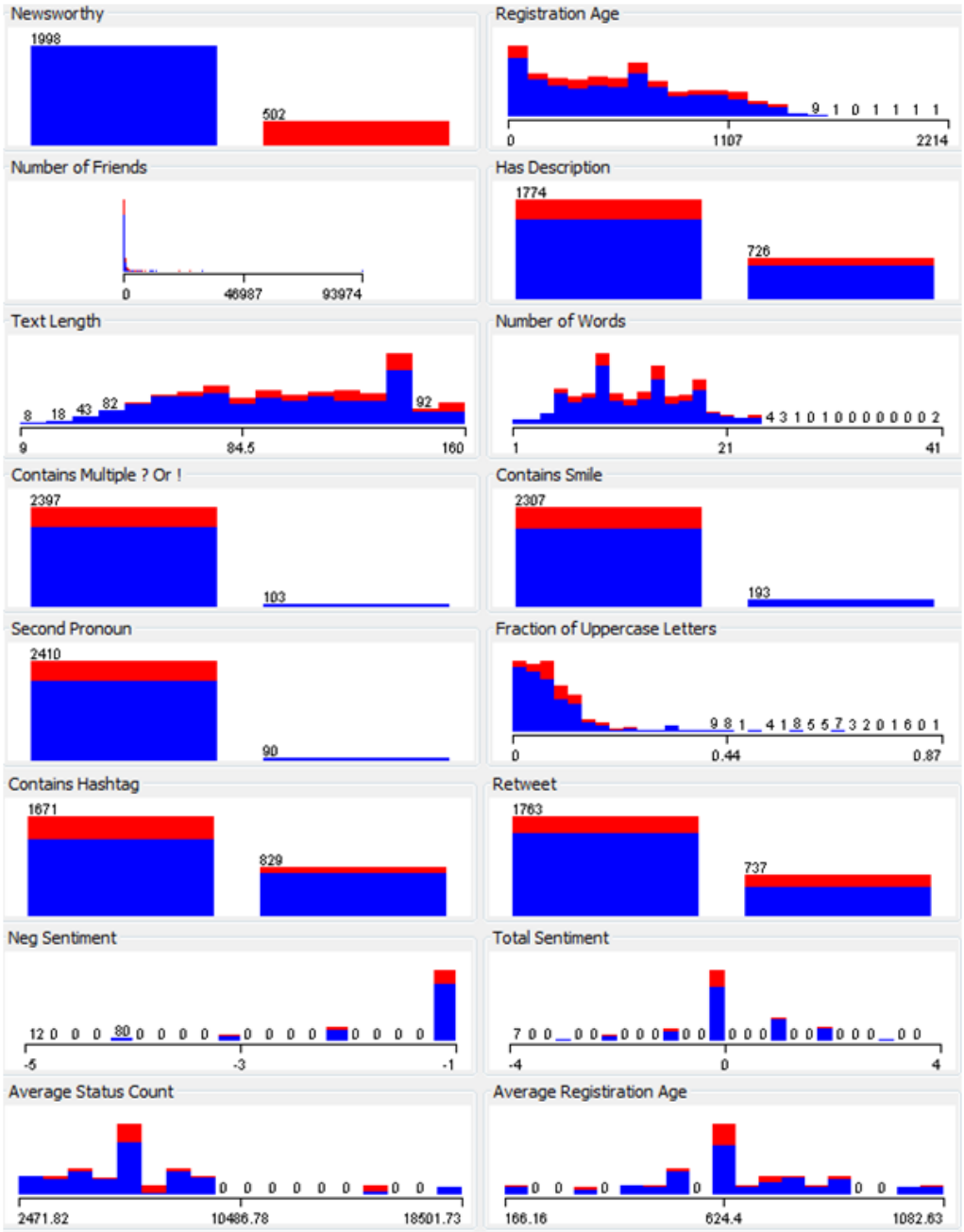

Figure A.1: Histogram of Atttibutes for Newsworthiness Label - I

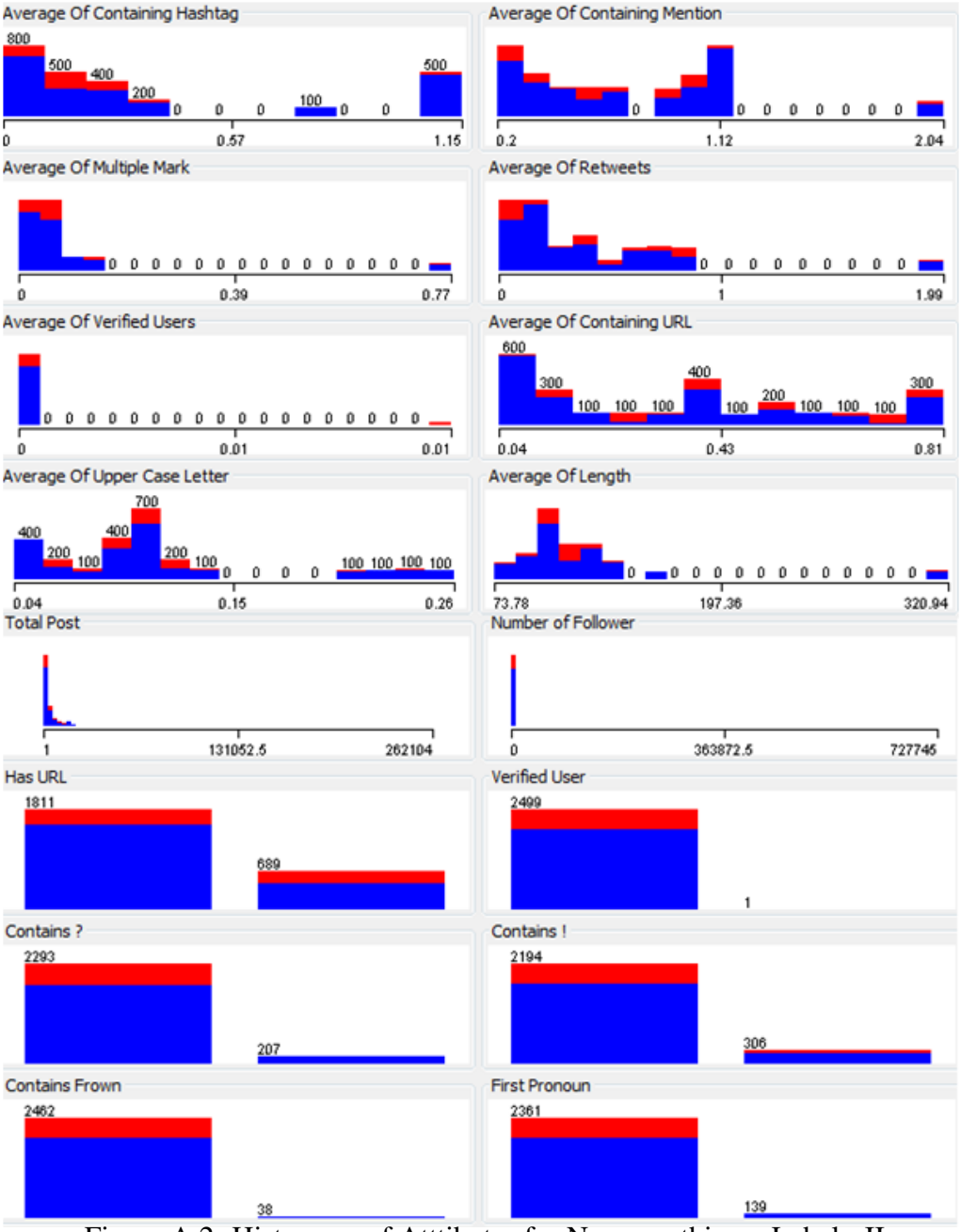

Figure A.2: Histogram of Atttibutes for Newsworthiness Label - II

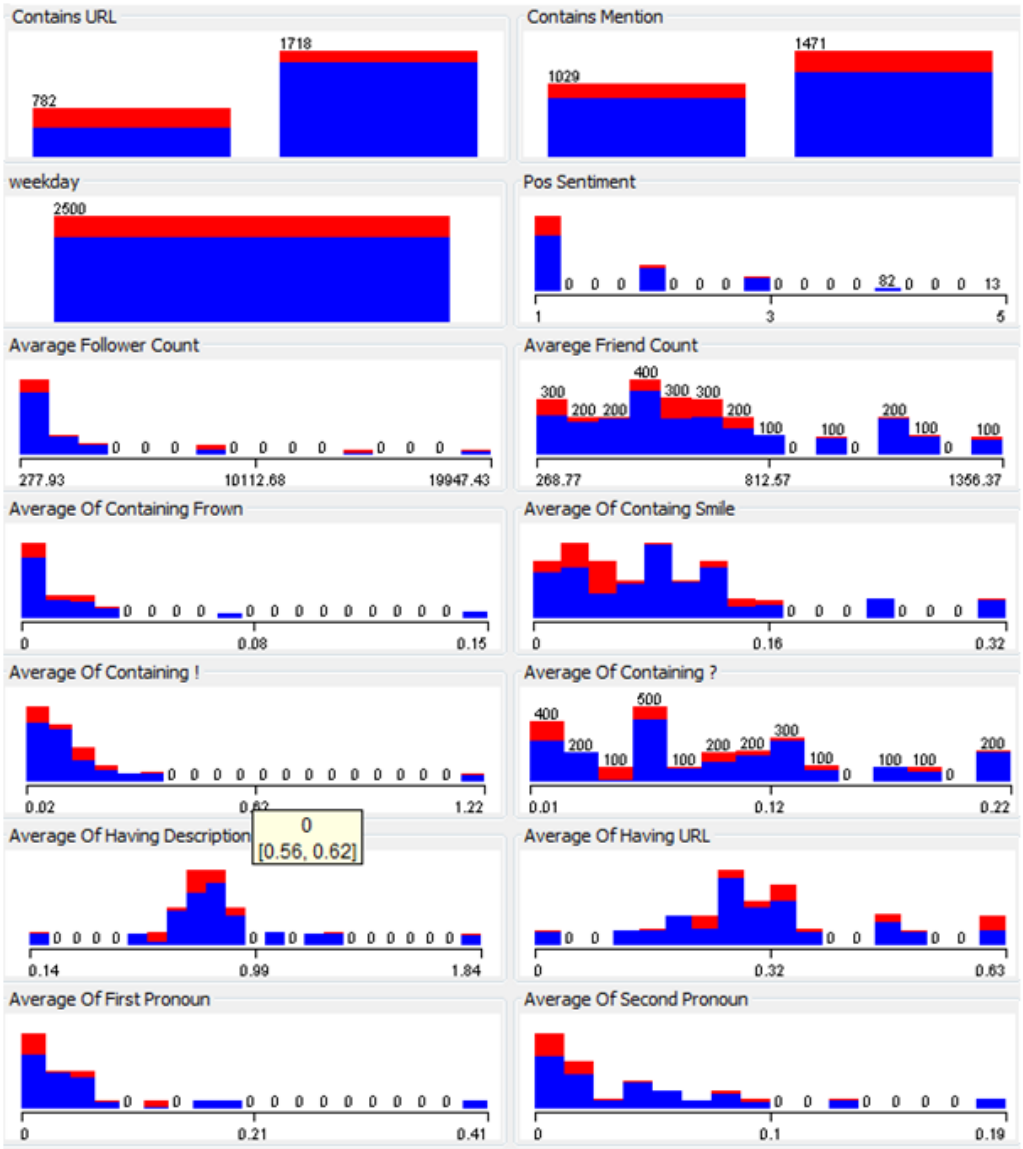

Figure A.3: Histogram of Atttibutes for Newsworthiness Label - III

APPENDIX B

HISTOGRAM OF IMPORTANCE LABEL

You can see histogram of importance label in Figure B.1, B.2 and B.3. Blue color represents important tweets and red color represents unimportant tweets.

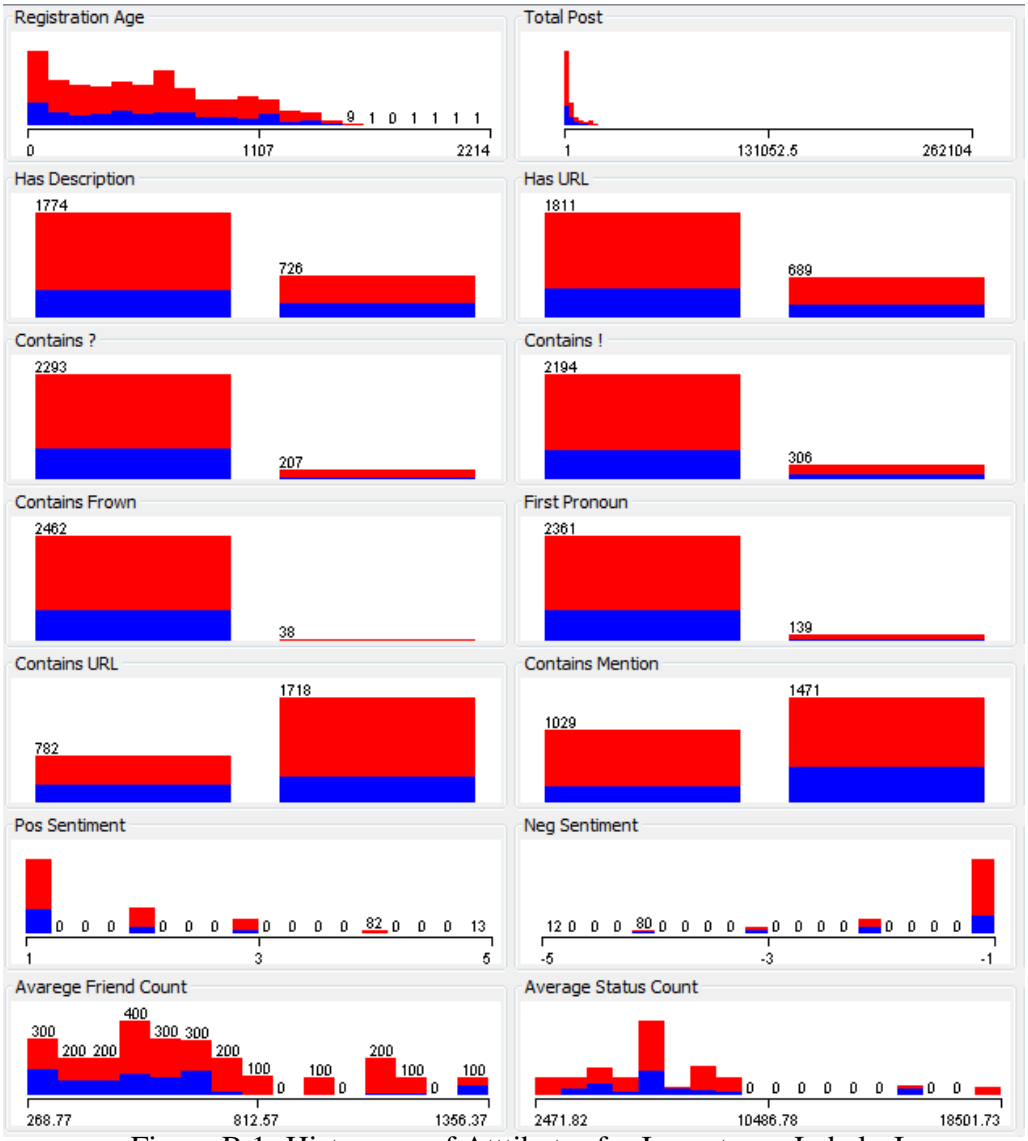

Figure B.1: Histogram of Atttibutes for Importance Label - I

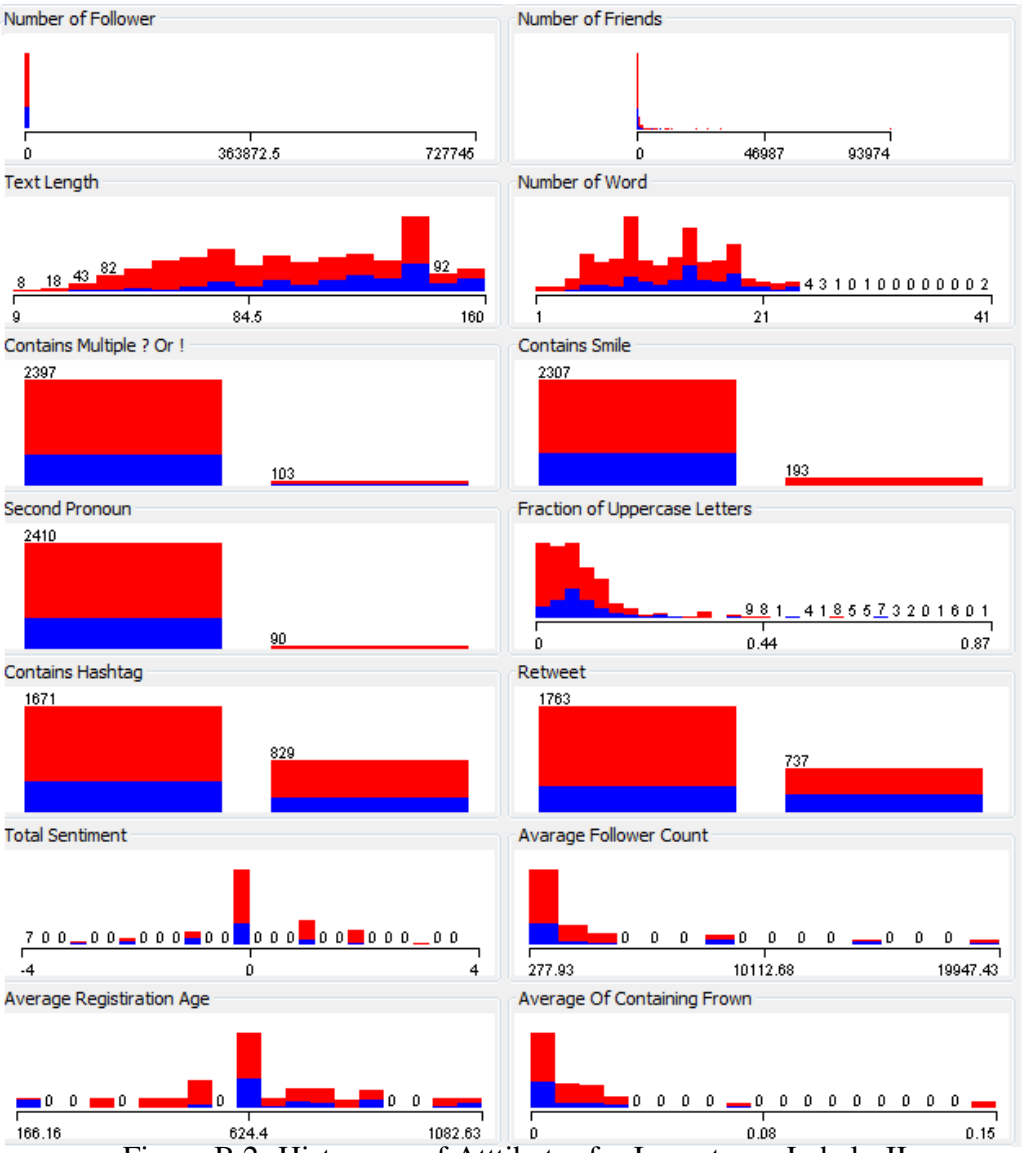

Figure B.2: Histogram of Atttibutes for Importance Label - II

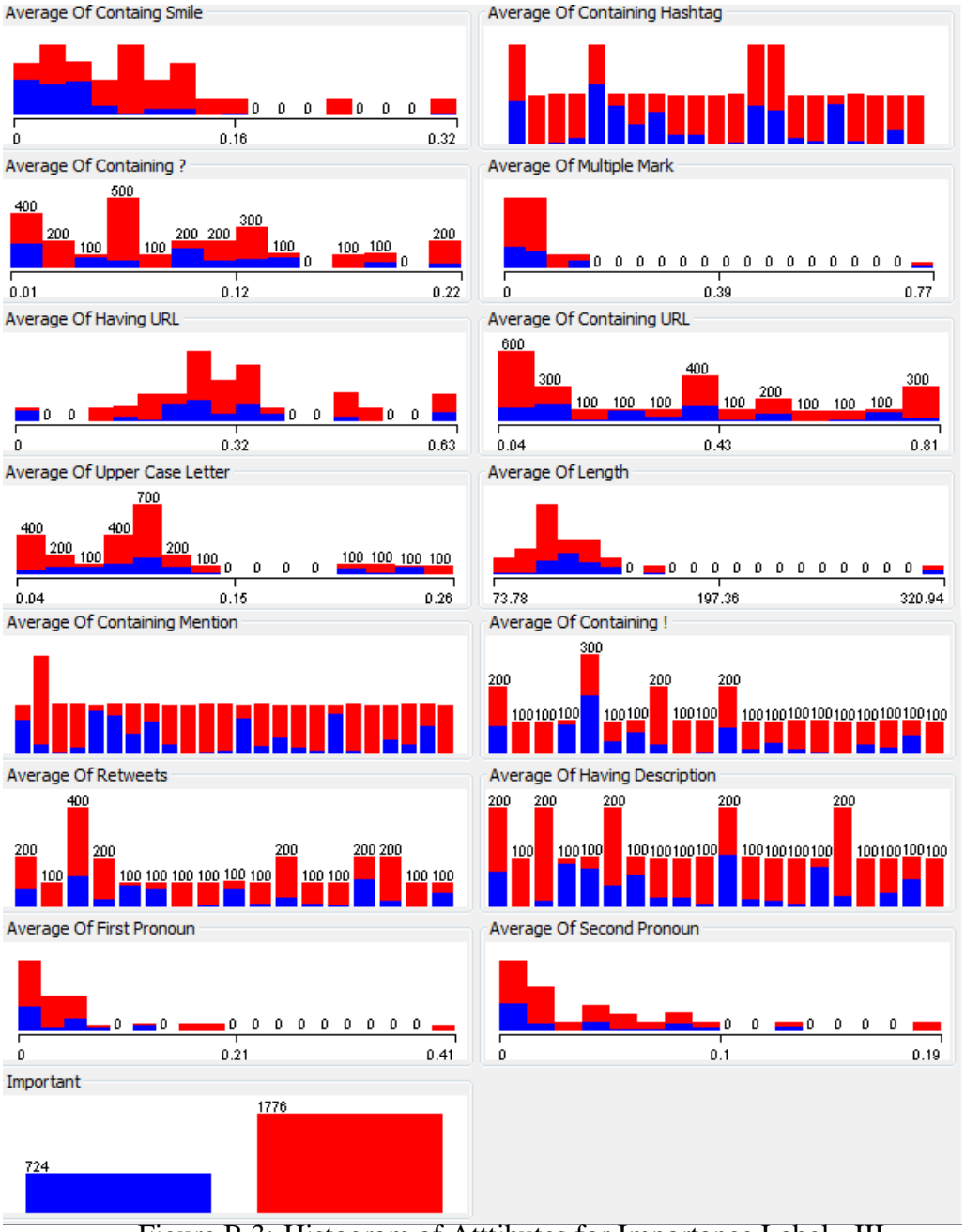

Figure B.3: Histogram of Atttibutes for Importance Label - III

APPENDIX C

HISTOGRAM OF CORRECTNESS LABEL

You can see histogram of correctness label in Figure C.1, C.2 and C.3. Blue color represents true tweets and red color represents false tweets.

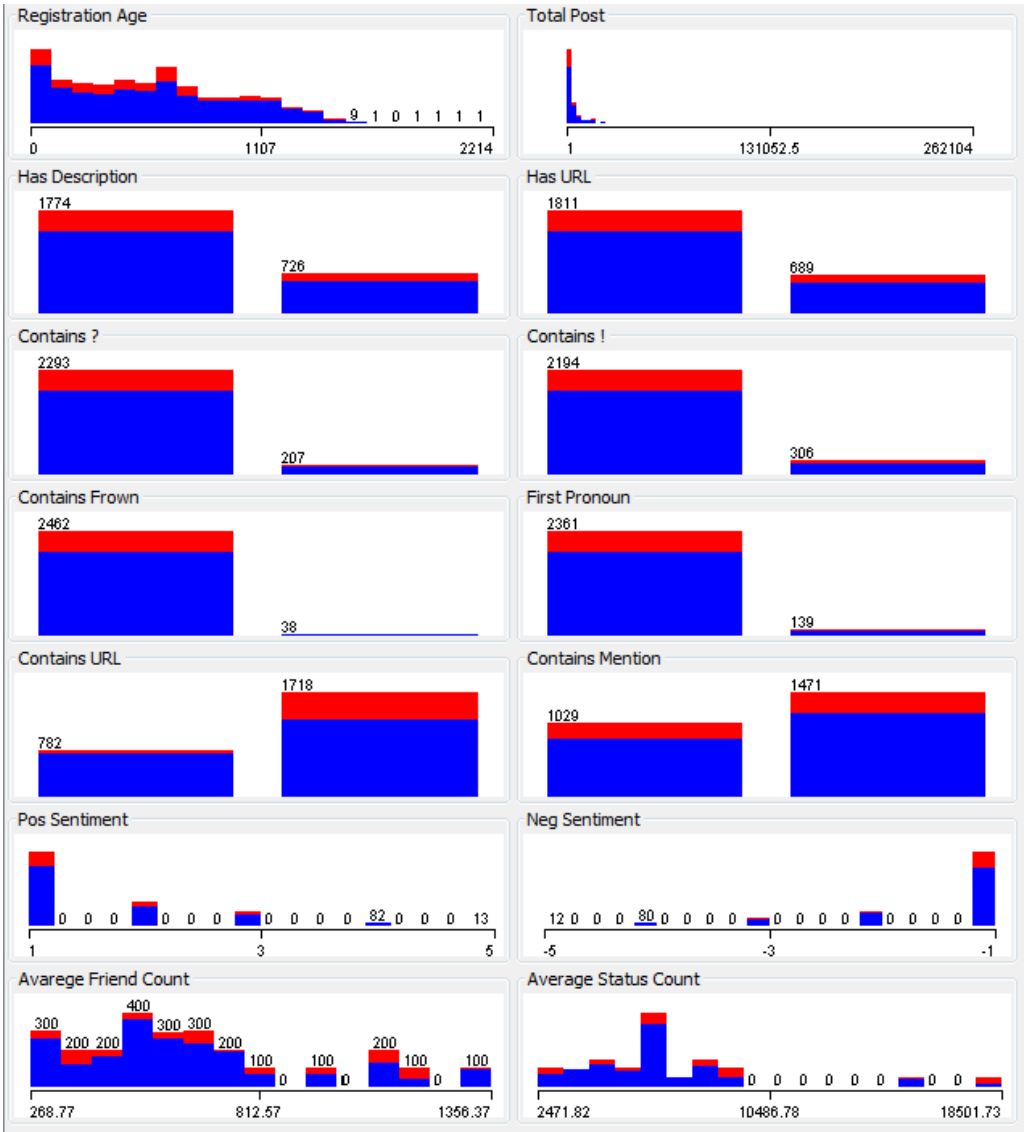

Figure C.1: Histogram of Atttibutes for Correctness Label - I

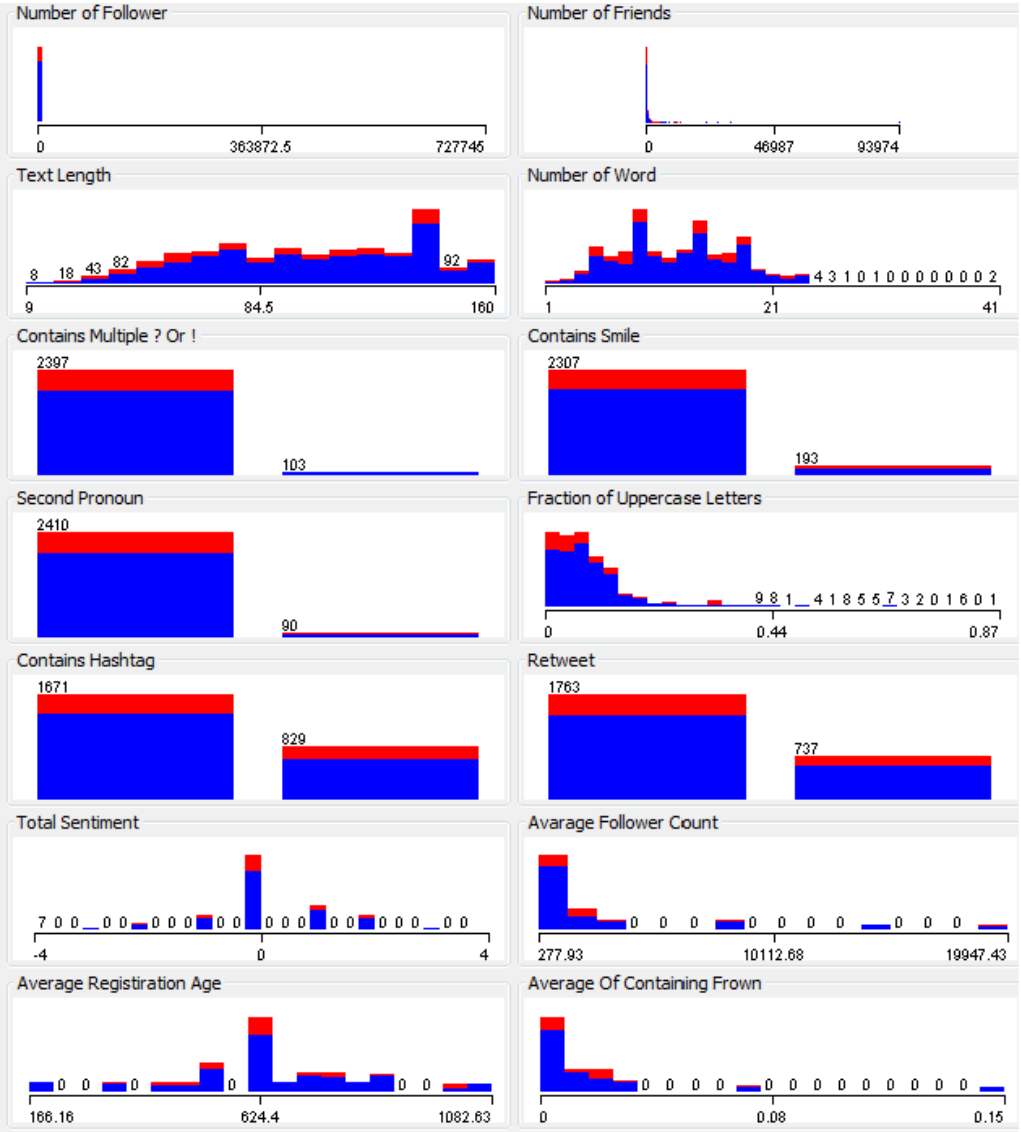

Figure C.2: Histogram of Atttibutes for Correctness Label - II

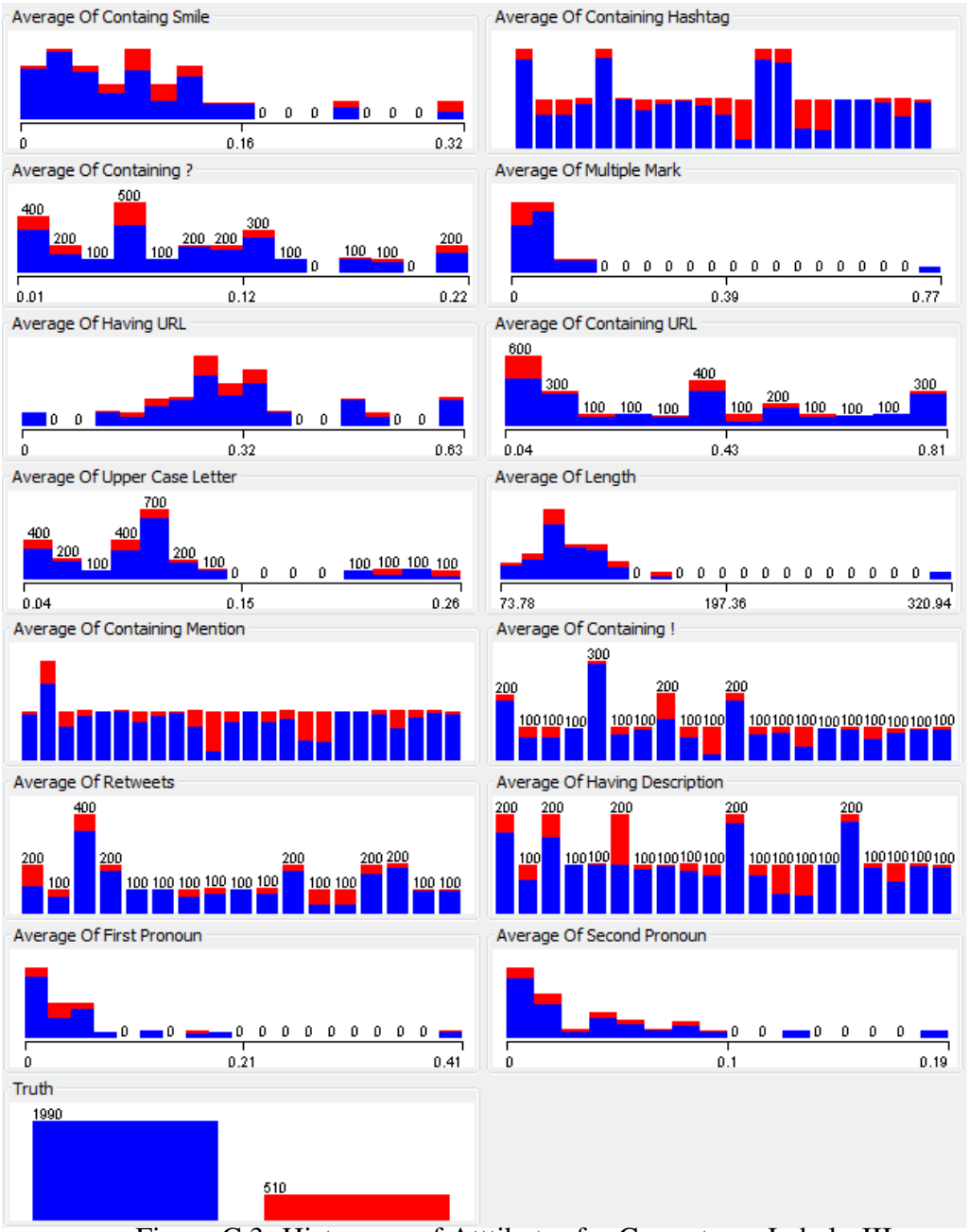

Figure C.3: Histogram of Atttibutes for Correctness Label - III