CREDIBILITY ANALYSIS ON TWEETS FOR NEWS AND DISCUSSION PROGRAMS BY USING A HYBRID CREDIBILITY ANALYSIS METHOD

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

ALI FATIH GÜNDÜZ

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

JANUARY 2016

Approval of the thesis:

CREDIBILITY ANALYSIS ON TWEETS FOR NEWS AND DISCUSSION PROGRAMS BY USING A HYBRID CREDIBILITY ANALYSIS METHOD

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

Prof. Dr. Gülbin Dural Ünver
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. Halit Oğuztüzün
Computer Engineering Department, METU _____

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

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

Assist. Prof. Dr. Selim Temizer
Computer Engineering Department, METU _____

Prof. Dr. İlyas Çiçekli
Computer Engineering Department, Hacettepe University _____

Date: 26.01.2016

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: ALI FATIH GÜNDÜZ

Signature :

ABSTRACT

CREDIBILITY ANALYSIS ON TWEETS FOR NEWS AND DISCUSSION PROGRAMS BY USING A HYBRID CREDIBILITY ANALYSIS METHOD

Gündüz, Ali Fatih
M.S., Department of Computer Engineering
Supervisor : Assoc. Prof. Dr. Pınar Karagöz

January 2016, 80 pages

In this work, we have studied credibility analysis of microblogging about news and discussion programs broadcast on television. We collected our data from one of the most important microblogging services, Twitter. Our credibility definition is based on three dimensions: being free from slang words, free from spamming purposes and newsworthy or important. We developed a hybrid model of supervised learning approach and graph based hub and authority score transferring approach. Firstly applying feature based classification on the collected data set and obtaining initial results, we tried to improve classification performance by graph based part of our study. Our graph based improvement approach is proposed to uncover the credibility relevance between microblogging messages and writers of those messages. We focused on message-message, message-writer and writer-writer connections in this graph. The performance of the proposed method is analyzed through a set of experiments. The final credibility score of a message is deduced based on each three dimension results at the end.

Keywords: Twitter, microblog, Credibility, Authority Transfer, television, news, discussion programs

ÖZ

HİBRİT BİR GÜVENİLİRLİK ANALİZİ METODU KULLANILARAK HABER VE TARTIŞMA PROGRAMLARI ÜZERİNE YAZILAN TWEETLERİN GÜVENİLİRLİK ANALİZİ

Gündüz, Ali Fatih
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Pınar Karagöz

Ocak 2016, 80 sayfa

Bu çalışmada teleziyonda yayınlanan haber ve tartışma programları hakkında yazılan microblogların güvenilirlik analizini yaptık. Verilerimizi en önemli mikroblog servislerinden biri olan Twitter'dan topladık. Güvenilirlik tanımımız üç boyuta dayanmaktadır: küfür kelimelerinden arı olmak, dikkat dağıtma amacında olmamak ve haber değeri taşımak veya önemli olmak. Gözetimli öğrenme ve karar ağacı yaklaşımlarının hibridi olan bir model geliştirdik. Öncelikle topladığımız veri seti üzerinde özelliklere dayanan sınıflandırma uygulayarak ilk sonuçları elde ederek karar ağacı kısmında sınıflandırma performansını yükseltmeye çalıştık. Grafiğe dayanan performance yükseltme yaklaşımımızı mikroblog mesajları ile bunların yazarları arasındaki güvenilirlik ilgisini çözmek için tasarladık. Bu grafikte mesaj-mesaj, mesaj-yazar ve yazar-yazar bağıntısına odaklandık. Önerdiğimiz metodun performansı deney setleriyle sınandı. Mesajların en son güvenilirlik skorları her üç boyuttaki sonuçlara dayandırılarak çıkarıldı.

Anahtar Kelimeler: Twitter, mikroblog, Güvenilirlik, Karar Ağacı, televizyon, haber, tartışma programları

To my family, my sisters and my fiancee

ACKNOWLEDGMENTS

First I would like to thank my family for their continuous support. They always encouraged me during my academical studies. Especially my father helped me a lot by sharing his knowledge and experience.

Moreover I am thankful to my supervisor Pınar Karagöz. She did not only teach me academical curriculum and guide my studies but also she was a graceful and kind person from whom I learned a lot.

Another person I thank is Dilek Önal for helping me in first stages of this study, especially during data gathering part.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

F1	F1 Score
NCP	No Class Precision
NCR	No Class Recall
YCP	Yes Class Precision
YCR	Yes Class Recall

CHAPTER 1

INTRODUCTION

Communication means provide a perfect research environment. From past to today, we have always observed change in the technology and the tools of communication. Word of mouth, newspapers, journals and other written means were once the only way to learn new information. However communication by radio and television inspired people by both audio and visual message transfers. Today current trend is communication through internet and mobile phone technologies.

Internet based social media tools are on the rise not only in Turkey as well as in the world. The number and variety of web-based media platforms available is impressive. Twitter is one of them. Twitter is a microblogging service that is being used by millions of users from all over the world. It allows users to post and exchange 140 character long messages, which are also known as tweets. Tweets can be published by sending emails, sending short text messages directly from smart phones and using a wide array of web-based services. Therefore Twitter facilitates real time propagation of information to a large group of users.

On the other hand, TV is still the most influential media resource in Turkey today. Many people learn social events and news from television programs. However TV is not the single mean of information any more. People nowadays do not only watch TV but also ask questions, make comments and discuss about the program content with their friends and families by using internet based communication tools such as Facebook updates, microblogging services, emailing and Twitter statuses. In this study we have focused on intersection of TV and Twitter[1] micro blogging communication

[1] https://twitter.com/

tools for Turkish news/discussion programs and microposts written for them.

Many TV programs ask and encourage their audience for participation. Most of them have Twitter accounts to enable their audience to contribute in the program flow by asking questions, making comments and expressing their feelings by writing tweets with program specific hashtags or mention tags. Therefore, hosts of those programs read those tweets and direct the program accordingly if they desire to do so.

However reading and classifying huge number of tweet messages during the program manually is not an easy task. Separating junk from useful information is a big challenge. Especially time is very limited in this case. Our aim is to study and develop a new and effective social media credibility analysis method based on data mining techniques, such that it can be utilized by TV programs to pick credible and useful postings while the program is on air.

However, it is a challenge to tell what is credible and how a message can be defined to be credible. Credible is defined as "able to be trusted or believed" by Cambridge dictionary[2]. By its nature, credibility is a subjective matter and it is always open to discussion. In addition its measurement depends on individual opinions and changes greatly for different people. Fogg and Tseng [11] state that credible information is believable information and they described credibility as a perceived quality composed of multiple dimensions.

In this study, we adapt meaning of credibility as being appropriate enough to be read during a TV program. We based our credibility definition on three dimensions: **being free from offensive words, being free from spamming and being newsworthy**. Our main objective is to determine if we can automatically assess the credibility of textual content of the posted tweets.

To achieve this task, we propose a hybrid solution. Both feature based supervised learning techniques and tweet-user, tweet-tweet and user-user interconnection structure based techniques are combined in our method. We examine both tweets and users to build a single clique graph for the graph based part of our study.

The contributions of this thesis work can be summarized as follows:

[2] http://dictionary.cambridge.org/dictionary/english/credible

- We brought a new credibility definition based on three dimensions: being free from offensive words, not being spam and being newsworthy.
- We applied a hybrid method of feature based and graph based approaches. Even though there are other hybrid studies, our study differs from them with respect to its feature set and graph creation phase. We build a connected graph in which user-user connections are created with friendship/followership relation, tweet-user connections are created with text/writer relation and tweet-tweet connections are created with contextual normalized similarity.
- In our study we used Turkish tweets as the data source. This study can be applied for other languages with some modifications as well. Turkish natural language processing tool component should be replaced with target language processing tool for this purpose.
- We created our data set from tweets written for current TV programs about social and political discussions.

Thesis organization is as follows: Chapter 2 consists of two parts. In the first part, especially focusing on Twitter we gave basic information about microblogs. In the second part we summarized our literature research on credibility analysis. Chapter 3 is dedicated to explanations of the proposed method. Chapter 4 presents experiment results of this study. Finally, Chapter 5 includes conclusions and future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

In this chapter, we give basic information about microblogs and Twitter. Then we summarize the studies that are closely related to our work.

2.1 General Information About Microblogs

Social network media became extremely popular for many reasons. Following personal and professional pursuits, building and sustaining friendship relations, advertising business promotions and many other purposes are met by those online microblogging services. Another reason as to why social networking attracted public attention is the ability to post spontaneous events easily. People express their feelings by writing about topics ranging from their daily and ordinary activities to cultural and social events by using their smart phones, tablets and other network connected devices with short sentences or images.

Microblog is a type of blog that lets users post short entries such as status updates, photos and liking comments on social networking websites. Microblogging is described by Wikipedia[1] as "a broadcast medium which differs from traditional blogging in that its content is typically smaller in size". In the recent decades we observed a rapid spread of microblogging websites and microblogging slowly moved into the mainstream. Several startups launched online microblogging environments like Twitter, Tumblr, FriendFeed, Plurk, Jaiku and identi.ca. Facebook, MySpace, LinkedIn, Google+ and XING are other famous websites that provide microblogging service.

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

This social media revolution found itself a place at research area as well. Kaplan and Haenlein [18] claim that creation of ambient awareness, unique form of push-pull communication and serving a virtual exhibitionism and voyeurism platform were specific characteristics of microblogs that paved the way for the aforementioned success. Scholars from different disciplines focused on various fields involving effects of microblogging on society. For example Tumasjan et al. [29] considering successful use of social media in the US presidential campaign of Barack Obama, investigated whether Twitter can be used to predict federal election results and understand political sentiment for coalitions in Germany.

Paul and Dredze [27] state that individual microblogging messages may contain little informational value alone but when millions of them are aggregated it becomes an important knowledge which can be used to gain overall health informatics of a society.

Sakaki et al. [28] propose a real-time earth quake alert system using Twitter microblogging messages in which they look for earth quake related keywords. As an application they constructed a probabilistic disaster detection solution to Japan's numerous earthquakes by using large number of Japanese Twitter users as disaster reporting sensors to inform people earlier.

In this study we used data collected from Twitter, therefore, in the rest of this section, we provide detailed information about Twitter so we spared rest of this section to detailed information about Twitter. Twitter is a key player in social media even their company name is used synonymously with microblogging today.

Twitter started its microblogging and online text message sharing service in 2006. Today it's a huge company with 302 million active users. Every day 500 million text messages (a.k.a. tweet) are shared online[2]. Twitter provides an online environment in which people can share their ideas, comments and concerns. Users of Twitter write tweets to update their statuses. Those written messages cannot exceed 140 character length. The users sometimes overcome this restriction by posting images of textual messages and URLs of information sources.

Every user in Twitter has to have a unique username. Users can read, favorite and

[2] https://about.twitter.com/tr/company

share tweets of other users which is called retweeting. Moreover users can create networks among themselves as following any other user or being followed by some other user as well. Users can send private direct messages to each other as well as writing public messages mentioning other user. Direct messages cannot be read from other Twitter users however the other tweet messages can be publicly accessed within Twitter.

To send a tweet in which users mention about each other, tweet should contain a specific character '@' in front of a username. This act is called as mentioning and this tag is called as mention tag. Similarly users can use another tag by adding '#' character in front of a specific word to create hashtags. Twitter gather the tweets containing same hashtag together so that users can share tweets about a specific discussion. Moreover Twitter promotes most active 10 discussion topics to its users to attract user's attention. Users can search hashtags, mention tags or any ordinary word via Twitter's facilities.

Twitter enables researchers to read, query and collect tweets of users who do not disable publicly visibility of their statuses. In this study we collected tweets written for current Turkish TV programs about politics, economics and general daily discussions.

2.2 Related Work

In the literature, there are several work conducted on the credibility of microblog messages. In this section, we present of a summary of these related studies.

In 2012, Kang et al. [17] proposed two definitions for tweet credibility as "degree of believability that can be assigned to a tweet about a target topic" and "expected believability imparted on a user as a result of their standing in the social network". Moreover they stated that [16] credibility is a function of perception consisting of the object being perceived and the person in 2015. Fogg [12] expressed website credibility in terms of prominence and interpretation which are defined as likelihood of being noticed and judgement of people noticed the element in his study.

Castillo and Yamaguchi studied both credibility assessment and newsworthiness of

tweets [8]. In their study, they focused on credibility of information and used the term credibility in the sense of believability. They classified tweets as credible or not. They randomly selected 383 topics from Twitter Monitor[3] [22] collection and get it evaluated by Mechanical Turk[4] by asking evaluators if they consider that a certain set of tweets as newsworthy or only informal conversations. Then they asked another group to read the text content and state if they believe that those tweets are likely to be true or false. In this evaluation they considered four levels of credibility and asked evaluators to provide justification in that fuzzy format. They proposed a supervised learning based method to automatically assess the credibility level of tweets which has a precision and recall rate between 70% and 80%.

O'Donovan et al. [24] underlined the fact that when studying credibility it is important to consider both the data type and methods used to generate ground truth. They proposed a method to simplify the feature space of credibility related dimensions and investigated features such as existence of URL, number of days the status has been on Twitter, number of followers of the user and sentiment score of the context etc. to predict most important feature dimensions of credibility. They also analyzed the effect of retweet chains and dyadic interpersonal communications written by '@' mention tag.

Detecting and preventing spam tweets is another aspect of credibility. Not only individuals write those tweets but also designed tweet generator tools are used to carry out this annoying and potentially malicious activity. Ferrara et al. [10] states that hundreds of thousands of social, economic and political incentives presented by highly crowded social media ecosystems attract spammers to design human imitating bot algorithms. Forelle et al. [13] states that bots are used for political lobbying in several countries like Russia, Mexico, China, UK, US and Turkey.

Twitter attaches importance to the fight against the spammers in order to sustain a spam-free social environment. They encourage [5] their users to report both profiles and individual tweets for spamming. Moreover they present technical solutions such as link shortener (t.co) to detect whether links lead to malicious contents as well.

[3] http://www.twittermonitor.net
[4] http://www.mturk.com/
[5] https://support.twitter.com/articles/64986?lang=en

To detect Twitter spam, there are two different approaches in the literature: focusing on the user classification and examining tweet content. In the first approach, profile details of the user, number of followers and friends, recent activities in the previous weeks, user behaviours and tweeting frequencies are investigated. Studies like [30], [6] and [32] aimed to classify users as spammers and non-spammers according to these user attributes.

The second approach considers topics of the tweets, duplications between the tweets, urls in the tweets, number of words and characters in the texts are searched. Martinez et al. [21] presented an example for this approach in which they detected spam tweets without any previous user information but by using contextual features obtained by natural language processing.

However as Yang et al. [32] expressed that Twitter spammers are developing counter strategies to evade detection as well. Tactics like purchasing followers, exchanging followers, mixing original tweets with spam content and using tools like Spinbot [6] to reduce duplication are developed by those malicious people to infiltrate the spam detection.

There are hybrid solutions of user based and content based approaches like [23] and [4] as well. Bara et al. [4] proposed a three step solution in which they firstly look for malicious links provided by Twitter database, secondly they look for pattern similarities between spam tweets and original tweets and finally they construct a bipartite network between users and corresponding tweets.

Clark et al. [9] proposed a solution to the problem of separating automated spam generators from human tweeters by a classification algorithm operating by using linguistic attributes like url count, average lexical dissimilarity and word introduction rate decay.

Kumar et al. [20] proposed a method to identify sources of information among active Twitter users during crises. They categorised users like generalists, specialist and information leaders in order to assess information quality coming from them during critical times.

[6] http://spinbot.com/about

Barbera [5] stated that ideologically similar users use same symbolic framework such as same type of language, similar average message length, same hashtags and retweeting similar tweets. Producing similar contents at Twitter opens the possibility to investigate how similar communities coexist in the online microblogging environment.

Alonso et al. [2] aimed to address the question whether it is possible to develop a strategy to predict interestingness of a tweet depending on a high quality labelled tweet train set. Ito et al. [15] proposed a method to assess tweet credibility using tweet-topic and user-topic features obtained from Latent Dirichlet Allocation model.

Pal et al. [26] studied this issue from another perspective by categorizing tweet writers. In their study, they tried to find most interesting and authoritative authors among millions of Twitter users for given specific topics. They compute self-similarity score for authors between their last two tweets so that they measure how similar an author writes tweets showing width of topics of interest. They also classify tweets into three categories: original tweets, conversational tweets and repeated tweets so that they consider the number of tweets in different categories of authors while deciding about their interestingness and clustering the users.

Abbasi and Liu [1] investigated credibility of the writer as well. They proposed CredRank algorithm to analyze users' online behavior to measure their credibility. They adopt credibility definition as "the quality of being trustworthy" and dividing it into three layers as message, source and media credibility, tried to rank social media users according to their credibility. They measured the behaviour similarity between users in order to cluster them if the similarity exceeds a given threshold value. In their study they did not only focus on Twitter but also other social media sources as well.

In addition to these learning based approaches, graph based solutions have been investigated as well. These solutions are basically use variations of well-known PageRank [25] and HITS [19] algorithms in the literature. Page and Brin, with PageRank, aim to measure and rate relative importance of Web pages mechanically. In this algorithm, the link structure among the web pages in the graph of web are considered. Being query independent and more sophisticated than simply counting links, PageRank ranks pages according to their importance of back links and forward links which directs to and are directed from the web page. With HITS algorithm, Kleinberg [19]

aimed to extract information from the link structure of network environment too. Although HITS is not solely specific to WWW, aiming to improve web search systems it identifies two kinds of web pages: authorities which are the pages that users look for to reach information and hubs which are pointer pages that lead to authorities. Kleinberg focused on the mutual relationship between those two kinds by giving non-negative invariant weights to each node and then making iterative score transfers between interlinked hub and authorities until scores converge to the equilibrium values.

In [14], Gun and Karagoz proposed a hybrid solution combining feature based and graph based methods for credibility analysis problem in microblogs. They chose three dimensions newsworthiness, importance and correctness to analyze credibility. Their study inspired our study and the techniques presented in [14] formed the basis for this thesis work. They focused on message, user and topic relationship and represented them in a graph structure linking each tweet with user and topic in the graph. They collected 43 features for those three kinds of nodes to use in feature based classification phase. Prediction results of first phase, number of followers of users and retweet numbers of tweets are mixed to assign initial scores to graph nodes. After they transfer scores in the graph, tweet nodes with final scores larger than predefined threshold are labeled as newsworthy, important and correct separately.

Another graph link structure based study is TURank which also constitutes a base to our study. Yamaguchi et al. [31] proposed Twitter user ranking algorithm (TURank) to determine authoritative users. They defined authoritative users as the ones who frequently submit useful information and they aimed to measure authoritativeness of users in order to rank them. They constructed a user-tweet schema graph where nodes are created from users and tweets and on the other hand edges are created from post, posted, follow, followed, retweet and retweeted relations between user-tweet, user-user and tweet-tweet nodes. Then they applied ObjectRank [3] on the user-tweet schema graph to evaluate the users' authority scores.

CHAPTER 3

PROPOSED METHOD

3.1 General Architecture

In this work, we examine credibility analysis from three dimensions: being free from slang words, being relevant (to the topic of the program) and being news-worthy. Each dimension is analyzed individually by our proposed method and at the end we propose an overall credibility measurement depending on those three dimensions.

Those three perspectives are:

1. Whether the tweet is offensive or contains slang words.
2. Considering the large number of online followers whether the tweet is written by spammers for irrelevant purposes like commercial advertisement, attracting people for some off-topic issue, distracting people by writing spam messages.
3. Whether the tweet is news-worthy/important/interesting.

We finally deduced that a tweet is credible only if it is free from slang words, is not a spam and news-worthy.

We collected tweets written for news and discussion programs broadcast at television. Those tweets are read by human volunteers and evaluated by them with respect to three perspectives of credibility. We collected their survey results to build our ground truth data base.

In this study, we propose a hybrid method consisting of two phases, which are feature

based supervised learning and graph based improvement phases. In the first phase, we determine the features to construct a supervised learning model and apply classification. In the second phase we aim to improve the obtained classification results by applying hub/authority score transfer iterations on a graph constructed from our data set. At the end of the iterations, each tweet gets its final authority score which is used to define the credibility of the tweet.

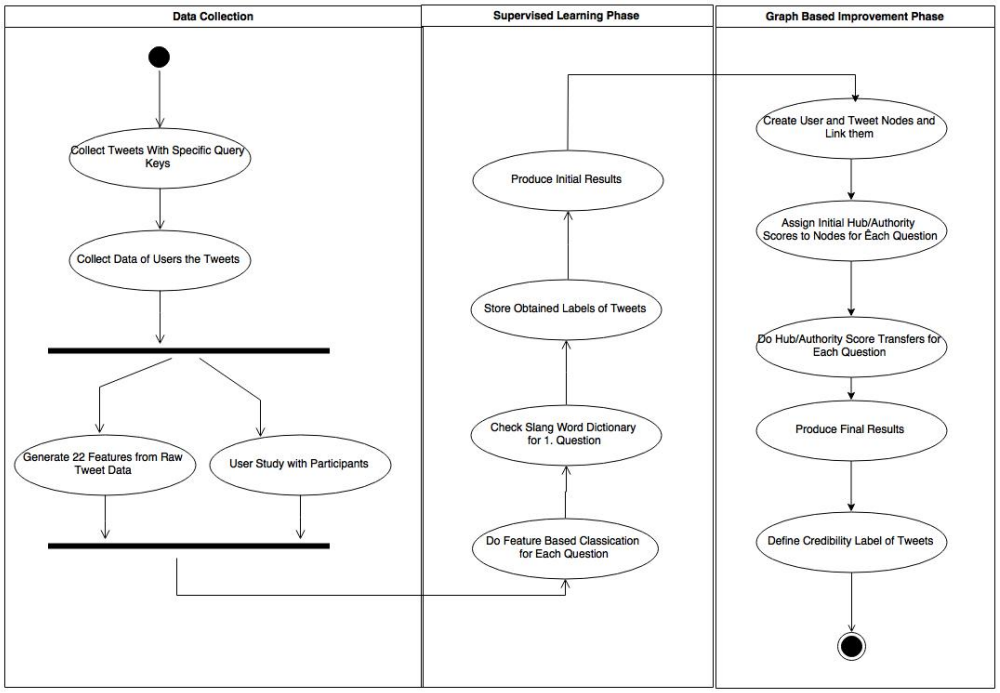

Figure 3.1: Activity diagram of the system

For analysing the first dimension of credibility, which is being free from slang words, we utilized a slightly different approach as well. We applied supervised learning phase, but skipped the graph based hub/authority score transfer phase. Instead we constructed a dictionary from slang words and checked the tweets whether they contain any word from the dictionary. Moreover, we made use of positive and negative sentiment score of tweet text as well.

Finally tweets are labelled as credible or not depending on the final results obtained for all the three dimensions at the end of graph based improvement phase. Our pro-

posed method is explained in detail in this section.

3.2 Tools And Libraries

In this study we used several tools that facilitated our work. Those tools and libraries are listed as follows:

1. Twitter4j [1]: Twitter data crawling part is handled by Twitter4j which is an unofficial Java library for the Twitter API.

2. SentiStrength [2]: SentiStrength is a sentiment analysis (opinion mining) API. It is free for academical research.

3. Weka [3]: Weka is a collection of machine learning algorithms for data mining tasks. In this study we used J48 Decision Tree algorithm of Weka API. Weka is open source software issued under the GNU General Public License.

4. Zemberek [4]: Zemberek, an open source Java library that provides morphological analysis and spell checking functions for Turkish and many other Turkic languages. We used this tool to obtain longest lemmas of words in text of tweets.

3.3 Data Collection

In this section, we describe how data set is gathered and ground truth is constructed in detail. Thirteen volunteers helped us to collect ground truth data. Totally 3000 tweets are crawled for this study and each tweet is read by 3 human voters via our web based ground truth collecting system. Each human voter is asked three yes/no questions about each tweet they read. Those questions are explained in 3.1 and they are used to create basis of our credibility analysis approach.

[1] http://twitter4j.org/en
[2] http://sentistrength.wlv.ac.uk/
[3] Weka, http://www.cs.waikato.ac.nz/ml/weka
[4] Zemberek Project, https://github.com ahmetaa/zemberek-nlp

3.3.1 Tweet And User Data

In order to build the data set, we selected 24 TV news/discussion programs and collected the tweets posted about these programs between 01.12.2014 and 13.12.2014. In our study, we crawled totally 3000 tweets by searching program-specific query keys. Also we crawled data of 1868 unique users all of which is a writer of at least one tweet in our tweet database. Those user data is used in the graph based phase to construct the user nodes of the graph network and this process will be explained in this chapter in detail.

Table 3.1 lists the channels, program names and broadcast times of the queried news and discussion programs during data collection. To obtain those data, we looked at broadcast schedules of all famous Turkish television channels. Among them, we only focused on the programs with political, social, economic and cultural discussion contents. Moreover we eliminated the ones with less than 20 tweets.

As it can be seen in the Table 3.1, all of the programs are broadcast between early night time and night time. Many of Turkish TV channels use this time zone to broadcast most interesting and important programs since many people watch TV at this time zone to learn about current social and political discussions.

Those TV programs have Twitter accounts and invite their audience to contribute to program flow by posting tweets written by using the specific query key or keys given in the table. Those query keys are generally mention tags and sometimes hashtags. Some of those query keys are twitter account mention tags of hosts of programs and we omitted those keys in order not to collect irrelevant tweets with the related ones.

The details of the statistics for the data set is given in Table 3.2

3.3.2 Constructing The Gold Standard For The Collected Data Set

In order to construct the gold standard for the evaluation, we conducted a user study with the contribution of 3 volunteers. In order to determine the label for each of the

Table3.1: Channels, program names and broadcast time

No	Channel	Program	Broadcast Time
1	A Haber	Kadraj	Mon 19:00, Tue 19:00, Thu 19:00
2	A Haber	%100 Siyaset	Mon 22:00
3	A Haber	Memleket Meselesi	Tue 22:00
4	A Haber	Birlikte Bakalım	Wed 22:00
5	A Haber	Deşifre	Fri 22:00
6	A Haber	Yaz Boz	Sat 22:00
7	Beyaz Tv	Son Söz	Tue 23:00
8	Beyaz Tv	Dinamit	Fri 23:00
9	CNN Türk	Ne Oluyor	Tue 21:00, Thu 21:00
10	CNN Türk	Tarafsız Bölge	Wed 21:00
11	CNN Türk	5N 1K	Sat 21:45
12	Habertürk	Türkiyenin Nabzı	Mon 20:00, Thu 20:00
13	Habertürk	Öteki Gündem	Tue 23:30, Thu 23:30, Sun 23:30
14	Habertürk	Karşıt Görüş	Wed 20:00
15	Habertürk	Dünyanın İşleri	Wed 23:30
16	Habertürk	Gündem Siyaset	Fri 20:30
17	Habertürk	Tarihin Arka Odası	Sat 23:15
18	Kanal 7	İskele Sancak	Fri 23:50
19	Kanal A	A Politik	Mon 21:30
20	NTV	Yakın Plan	Wed 21:10
21	NTV	Gündem Masası	Tue 21:10
22	Ulusal Kanal	Çıkış Yolu	Mon 21:00
23	Ulusal Kanal	Ceviz Kabuğu	Sat 21:00
24	Ülke Tv	Bıçak Sırtı	Mon 23:40, Tue 23:40

Table3.2: Program related tweet details

No	Program Name	Total Related Tweet	Number of Users Tweeted	Average Number of Tweets per Users
1	Deşifre	370	291	1,27
2	5N 1K	272	178	1,53
3	Son Söz	218	119	1,83
4	Tarafsız Bölge	213	163	1,31
5	%100 Siyaset	176	109	1,61
6	Oteki Gündem	174	119	1,46
7	Karşıt Görüş	165	105	1,57
8	Yaz Boz	154	97	1,59
9	Tarihin Arka Odası	147	115	1,28
10	Türkiyenin Nabzı	140	112	1,25
11	Dinamit	136	85	1,60
12	Kadraj	126	98	1,29
13	Memleket Meselesi	100	61	1,64
14	Ne Oluyor	76	48	1,58
15	Bıçak Sırtı	76	56	1,36
16	Ceviz Kabuğu	74	48	1,54
17	Yakın Plan	72	35	2,06
18	Çıkış Yolu	72	47	1,53
19	Birlikte Bakalım	69	50	1,38

three dimensions, we asked the following questions to the users [5]

1. Does the tweet contain swearing, abusing or offensive words?

2. Is the tweet written for distracting, unrelated, advertising or out of program scope purposes?

3. Is the content interesting, important or news-worthy?

The volunteers answered each of these questions as either Yes or No. The ground truth label is determined by using majority voting.

[5] As our tweet data were constructed from Turkish tweets, the questions above were Turkish in our website and volunteers were native Turkish speakers. Original questions in Turkish were:

1. Küfür, Hakaret, Saldırgan veya İncitici İfade İçeriyor mu?
2. Dikkat Dağıtıcı, Alakasız, Reklam İçerikli veya Program Dışı Bir Amaçla mı Yazılmış?
3. İçerik İlginç, Dikkate Değer veya Haber Değeri Taşıyor mu?

Table3.3: Raw tweet features and explanations of those features

Featuer Number	Feature	Explanation
1	Tweet Id	Unique id of tweet given by Twitter
2	User Id	Unique id of user given by the Twitter
3	Text	Textual body of the tweet
4	Retweet count	Number of times the tweet is retweeted
5	Is Retweet	Whether the tweet is a retweet
6	Favorite count	Number of times the tweet is favourited

3.4 Supervised Learning Phase

As explained in section 2, we inspired from the study of Gun and Karagoz [14]. In their study they used both Weka[6] and KNIME[7] classification tools and stated that they obtained better results with WEKA tool. They made experiments with 8 different classification algorithms and among them J48 decision tree classification algorithm appeared to be very promising. Since this part of our study is very similar to their study with respect to supervised learning approaches, we decided to use J48 decision tree classifier of Weka API for the supervised learning phase of our hybrid study.

We crawled user data from Twitter which is shown at Table 3.4. Those features are used in the graph based phase of our study which is explained in Section 3.5 in detail. In addition we crawled the raw tweet data from Twitter and we obtained the initial features displayed in Table 3.3. Those features are used to create the feature dimensions of supervised classification phase and their explanations are given in the Table 3.5. Totally we used 30 features, where 22 of them are tweet features and 8 of them are user features.

3.5 Graph Based Improvement Phase

3.5.1 Graph Construction

After applying the supervised learning phase of the proposed method, we aim to improve classification results with graph based improvement phase. Firstly we create

[6] Weka. http://www.cs.waikato.ac.nz/ml/weka
[7] Knime. http://www.knime.org/

Table3.4: Features of users and their explanations

Feature Number	Feature	Explanation
1	User Id	Unique id of the user given by Twitter
2	Friends count	Number of users followed by the user
3	Followers count	Number of user following the user
4	Favorites count	Number of tweets favorited by the user
5	Tweets count	Number of tweets written by the user
6	Listed count	Number of public lists the user is listed on
7	Friends list	List of ids of friends of the user
8	Followers list	List of ids of the followers of the user

a graph from the collected tweets and users' data, where each tweet and each user is represented as a node in this graph. As explained previously, we have 3000 tweets and 1868 users so totally our graph has 4868 nodes and it is an undirected graph. Links are created according to the following rules:

1. A user is directly linked to a tweet if he/she is the writer of the tweet.
2. A user is directly linked to a user if he/she is a follower/friend of that user.
3. A tweet is directly linked to a tweet if the tweet's content has equal to or more than a predefined cosine similarity with the other tweet's content.

In this phase of the study, we aimed to examine the effect of user-user network and tweet-tweet similarity factor in credibility problem. Our basic motivation is to investigate whether similar tweets are more likely to be credible together or not and similarly whether similar users are more likely to write credible tweets together or not.

In order to link the tweets, we need to find tweet similarities. To this aim, we first parse the text of the tweets and obtain the word sets and eliminate the effect of stop words. Those word sets are processed with Zemberek Turkish NLP tool and we replace them with their corresponding longest lemma term so that we could identify relations among the same words in different morphological forms. This textual data is converted to term vector for each tweet.

Term vector of a tweet contains longest lemmas of all unique words existing in its text and corresponding term frequency-inverse document frequency multiplication score

Table3.5: Features of tweets used at classification process of supervised learning phase and their explanations

Feature No	Feature	Explanation
1	Length of tweet	Number of characters in the tweet text
2	Fraction of upper case letters	Division of number of upper case characters with lower case characters of the tweet
3	Total number of words	Total number of words separated by spaces in the tweet
4	Number of words with mention tags	Such as @userName
5	Number of words with hashtags	Such as #topic
6	Number of words without @ and # tags	Number of words without '@' and '#' characters
7	Fractionof tagged words	Division of sum of features 4 and 5 by feature 6
8	Contains question mark	Whether the tweet text contains '?' character
9	Contains exclamation mark	Whether the tweet text contains '!' character
10	Contains smile emoticon	Whether the tweet text contains smile emoticons such as :-)
11	Contains frown emoticon	Whether the tweet text contains frown emoticons such as :-(
12	Contains URL	Whether the tweet text contains any form of url
13	Positive sentiment score	SentiStrength library based positive sentiment score of the text of the tweet
14	Negative sentiment score	SentiStrength library based negative sentiment score of the text of the tweet
15	Contains plural/singular first pronoun	Whether the tweet text contains "ben", "biz", "bana" etc.
16	Contains second plural/singular pronoun	Whether the tweet text contains "sen", "siz", "sana" etc.
17	Contains demonstrative pronoun	Whether the tweet text contains "bu", "şu", "o", "bunlar", "şunlar", etc.
18	Contains interrogative pronoun	Whether the tweet text contains "ne", "kim", "nerede", "nereye", "hangi", "kaç", etc.
19	Retweet count	Number of times the tweet is retweeted
20	Is retweet	Whether the tweet itself is a retweet
21	Favorite count	Number of times the tweet favorited by Twitter users
22	Is reply to a user	Whether the tweet is written to reply/direct to another user

pairs.

In order to obtain multiplication results firstly we calculated the term frequencies of the longest lemma terms of tweets according to Equation 3.1.

$$TermFrequency(T_i, w_j) = \frac{Number\ of\ times\ w_j\ occurs\ in\ the\ text\ of\ T_i}{Number\ of\ words\ in\ the\ text\ of\ T_i} \tag{3.1}$$

Then inverse document frequencies of words are calculated according to Equation 3.2.

$$IDF(w, D) = log_{10}(\frac{Number\ of\ Term\ Vectors\ (i.e.\ Number\ of\ Tweets)}{Number\ of\ Term\ Vectors\ Containing\ Word\ w}) \tag{3.2}$$

Those terms and their corresponding term frequency-inverse document frequency multiplication result pairs are used to obtain Tf-Idf based term vectors of tweets according to Equation 3.3.

$$Tf Idf\ Based\ Term\ Vector\ of\ T_i =< (w_1, tfidf_1), (w_2, tfidf_2), ..., (w_n, tfidf_n) > \tag{3.3}$$

Finally, for each tweet, we calculate cosine similarity of its term-vector with all others according to Equation 3.4. Depending on the cosine similarity measure, we link associated tweet nodes in the graph.

$$Cosine\ similarity\ between\ Tweet_i\ and\ Tweet_j = \frac{Tweet_i\ .\ Tweet_j}{\|Tweet_i\| * \|Tweet_j\|} \tag{3.4}$$

We wanted to construct a connected graph in which each node is linked to another node by some path, for this purpose we experimentally searched for the optimal cosine similarity threshold value to link two tweets.

As it is shown in Table 3.6, 0.063 cosine similarity threshold between tweet-tweet linking procedures appeared to be maximum threshold that enable constructing the

Table3.6: Tweet-tweet linking cosine similarity thresholds and corresponding graph structure

Cosine Similarity Threshold	Number of Cliques
0.25	85
0.20	44
0.15	5
0.10	2
0.09	2
0.08	2
0.07	2
0.069	2
0.068	2
0.067	2
0.066	2
0.065	2
0.064	2
0.063	1

desired graph. So we create tweet-tweet links in the final graph by using this threshold.

3.5.2 Random Walk Iterations On The Graph

After constructing the graph and providing initial hub/authority scores to the nodes, we run a predefined number of iterations in the graph for hub/authority transfers between nodes. Algorithms such as PageRank [7] and HITS [19] inspired us in this score distribution part. At the end of those iterations, a tweet is classified as positive if its final authority score is greater than zero, and classified as negative otherwise.

After assigning initial scores, we made iterations in the graph. As explained in Chapter 4, we made experiments with 1 to 3 iterations. During those iterations node hub scores are updated by adding a predefined ratio of authority scores of nodes linked to the node. Similarly authority scores are updated by adding a predefined ratio of hub scores of nodes linked to the node. Depending on the link structure of nodes,

hub/authority scores increased or decreased during those iterations.

$$N_j\ hub\ score = \sum_{i}^{set\ of\ linked\ nodes\ with\ N_j} weight * N_i\ authority\ score \tag{3.5}$$

$$N_j\ authority\ score = \sum_{i}^{set\ of\ linked\ nodes\ with\ N_j} weight * N_i\ hub\ score \tag{3.6}$$

According to the link structure, node hub and authority scores increases or decreases since scores are assigned positive or negative to the classes. Final authority score is checked and if it is greater than zero, the node is classified as positive class and vice versa.

3.6 Slang Word Analysis Approach

Unlike other two question, we tried a different approach than graph-based random walk iterations for slang word existence analysis. To classify tweets as positive or negative from the first dimension, we created a dictionary from the slang words in our database and checked existence of any of those words in the tweet text.

This dictionary is created from known slang words used in those tweet texts. We searched all tweets and recorded identified slang words in this dictionary.

However this method lacks the ability to detect true positives which are not written by using any slang word but metaphors and playing with words.

To overcome this problem, we made experiments with positive and negative sentiment scores of the tweet text. In addition we considered the effect classification results of first phase. Depending on those scores and initial feature-based classification results, we made positive/negative classifications. Those classification results of our experiments are explained in 4 in detail.

3.7 Overall Credibility Determination

In this study we accepted credibility as being free from offensive words, being dedicated to the purpose of the discussion and being newsworthy. Aiming to pick up the tweets appropriate enough to be read during a TV program, we determined overall credibility of a tweet based upon those three dimension. A tweet is classified as credible only if,

- It does not contain slang words and is not written in an offensive manner,
- It does not contain irrelevant information, advertisement and spamming, and
- It does conveys important, news-worthy or interesting information.

Any tweet lacking any of those three requirements is not classified as credible.

CHAPTER 4

EXPERIMENT RESULTS

In this chapter we explain and discuss the results of the experiments. Each dimension of the credibility problem is experimented individually and results are shown in bar charts in this section. In addition those results are given in tables in appendices in the corresponding sections as well.

Our data set consists of 3000 tweets and we applied ten-fold cross validation method in the experiments.

We assign different initial scores to test and train groups in the experiments. During experiments, training set tweet nodes are assigned their initial scores depending on their real classes. Members of positive class are given 1000 hub and authority scores. On the other hand, members of negative class are given -1000 * k hub and authority scores.

The k constant above is the ratio of the size of positive class to the size of negative class. Using k constant enabled us to stop a larger class dominating score transfer procedure on its behalf.

For example in a 2000 positive, 1000 negative tweet set; positive tweets would get 1000 and negative tweets get -2000 initial hub/authority scores during initialization. In the score transfer phase each tweet in the graph makes contribution to the tweets which are linked with it. Giving larger absolute value initial scores to negative tweet nodes protects them from being dominated by positive tweet nodes in score transfer phase since there are two positive tweets contributing to one negative tweet when we considered overall graph network.

On the other hand, test set tweets are given initial hub and authority scores depending on their expected classes according to the feature based classification phase. Positive classified tweet nodes are given 1000 hub and authority score and negative classified tweet nodes are given -1000 * k hub and authority scores.

Moreover user nodes are assigned initial score as well. We made three kinds of experiments with initial score assignment task for user nodes. Firstly, user nodes are given 0 initial hub and authority scores while tweet nodes are assigned initial scores as explained above. Secondly, we made experiments with 0 scored tweets and positive/negative scored user nodes. Finally, we combined best tweet and user scoring techniques to see the hybrid overall results.

While assigning initial scores to the user nodes, each user with true positive answers are assigned positive initial scores and each user without true positive answers are assigned negative initial scores. Similar to tweet scoring, positively classified nodes are given 1000 hub and authority score and negatively classified nodes are given -1000 * k hub and authority scores to avoid domination of larger class.

All final results of the experiments are given as the average scores of the 10-fold cross validation. We showed precision, recall and f1 scores in this section. In addition to those data displayed in the tables, true positive count, true negative count, accuracy, specificity and sensitivity values are given in appendices.

F1 scores, accuracy, specificity and sensitivity values are calculated according to the equations (4.1), (4.2), (4.3) and (4.4) below:

$$F1\ score = \frac{2 * precision * recall}{precision + recall} \tag{4.1}$$

$$Accuracy = (True\_Positive + True\_Negative)/Size\_of\_Data\_Set \tag{4.2}$$

$$Specificity : True\_Negative/(True\_Negative + False\_Positive) \tag{4.3}$$

$$Sensitivity = True\_Positive/(True\_Positive + False\_Negative) \quad (4.4)$$

4.1 Experimental Analysis For Dimension 1 - Slang Language

4.1.1 Only Tweet Initial Scoring Results

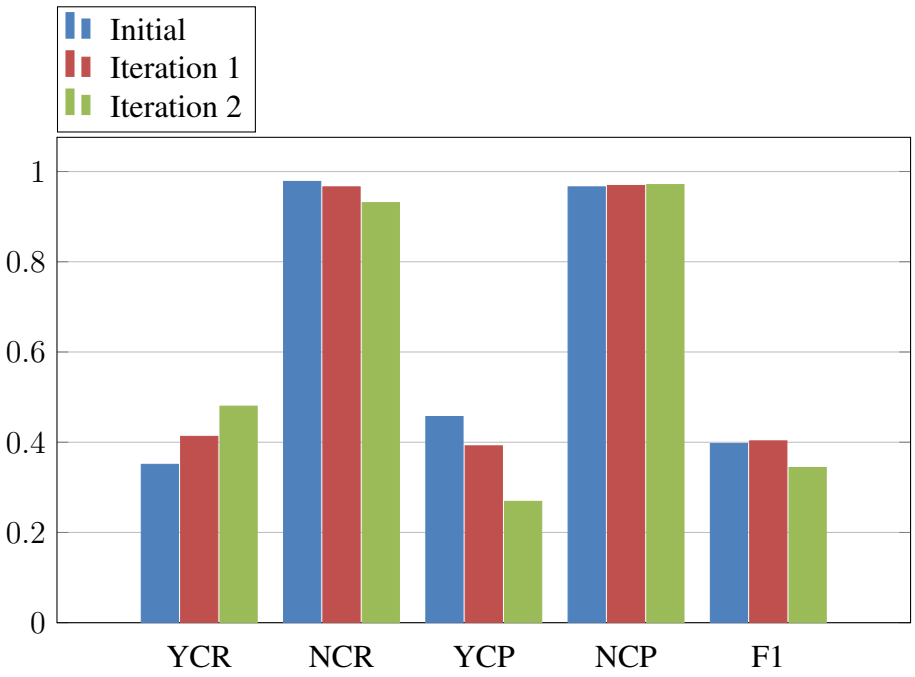

Figure 4.1: Random walk iterations with hub weight 0.01 and authority weight 0.01

In Figure 4.1 Initial and Iteration 2 experiments showed 36% increase in the Yes Class Recall(YCR) however Yes Class Precision(YCP) decreased 41%. No Class Recall(NCR) and No Class Precision(NCP) slightly changed. F1 score results decreased by 13%.

In Figure 4.2 YCR increased to 98% in the Iteration 2 experiment when we compare with initial experiment. However YCP decreased 50% and F1 score slightly decreased.

In Figure 4.3 between Initial and Iteration 2 experiments we see 83% increase in

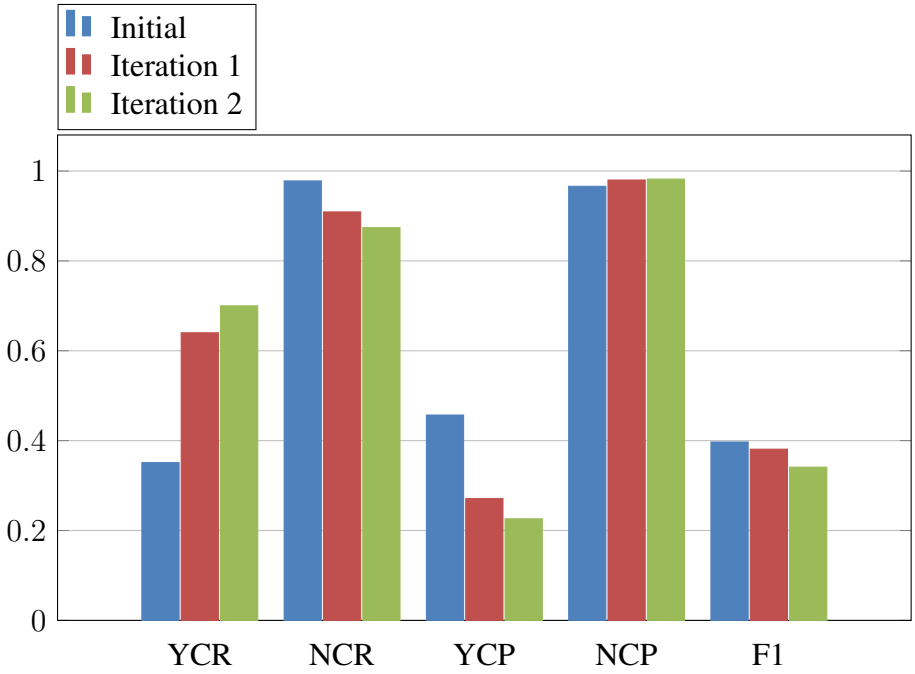

Figure 4.2: Random walk iterations with hub weight 0.04 and authority weight 0.04

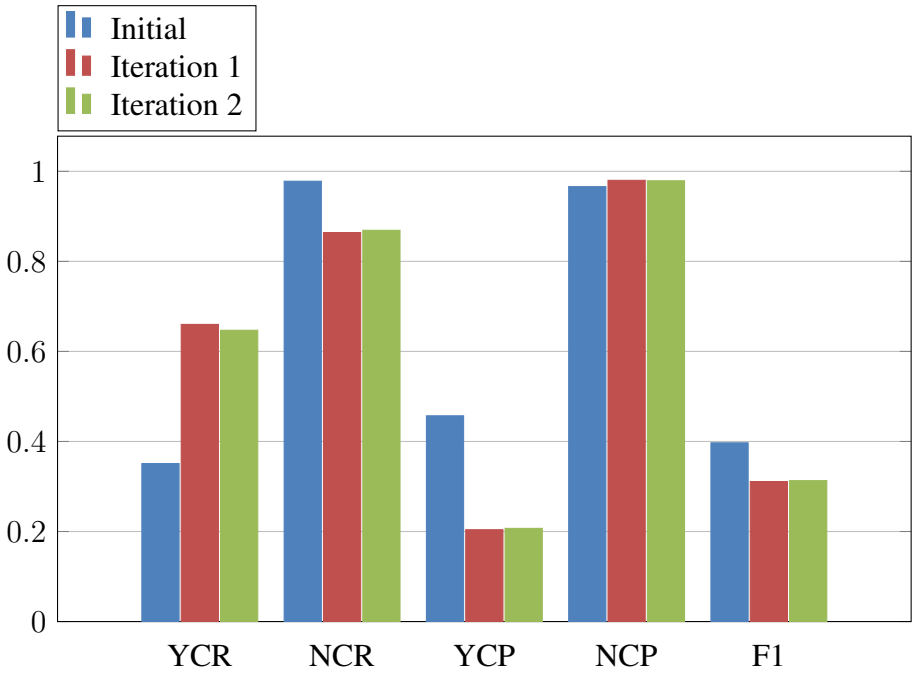

Figure 4.3: Random walk iterations with hub weight 0.07 and authority weight 0.07

YCR, 54% decrease in YCP, 11% decrease in NCR and 25% decrease in F1 score.

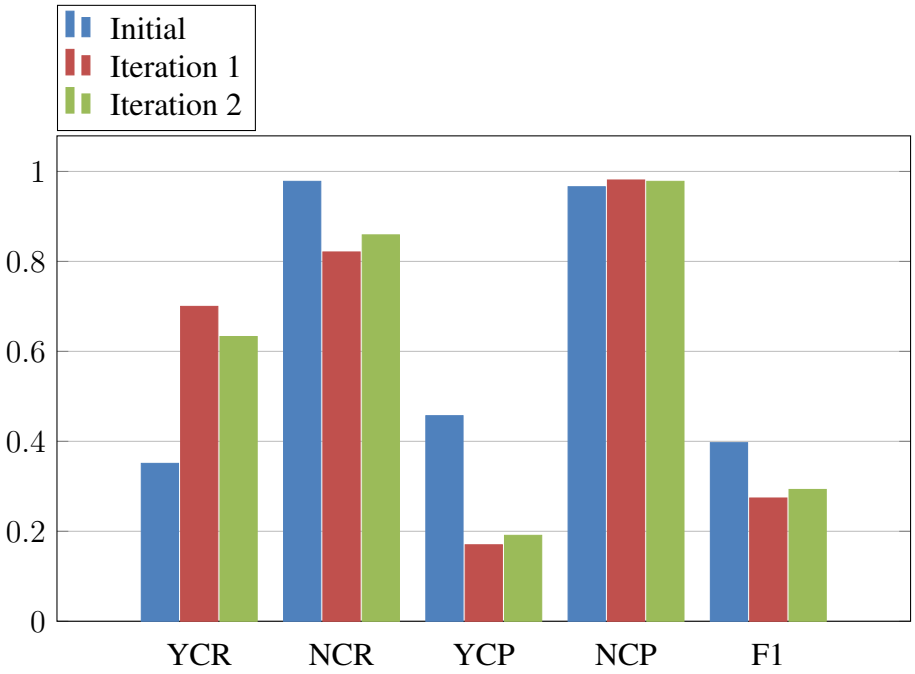

Figure 4.4: Random walk iterations with hub weight 0.1 and authority weight 0.1

In Figure 4.4 we see that the best YCR is obtained at Iteration 1 experiment which is 43% better than Initial experiment. YCP scores decreased in experiments Iteration 2 and 1. Iteration 2 F1 score appeared to be better than Iteration 1 but it showed 26% decrease when we compared Initial F1 score.

We observed that lower hub/authority weights led to better results. Best F1 score is seen at hub/authority weight 0.01 in the Iteration 1 experiment. Moreover this experiment has the best iteration based F1 score result among other first dimension related experiments as well. As weights changed from 0.01 to 0.1, we observed decrease in the F1 scores of experiments due to effect of larger weights. Initial graph structure represented better classification scheme which can only be improved by 0.01 hub and authority weight transfer experiment. The initial classification scheme is deteriorated as larger hub and authority scores transferred among the nodes of the graph.

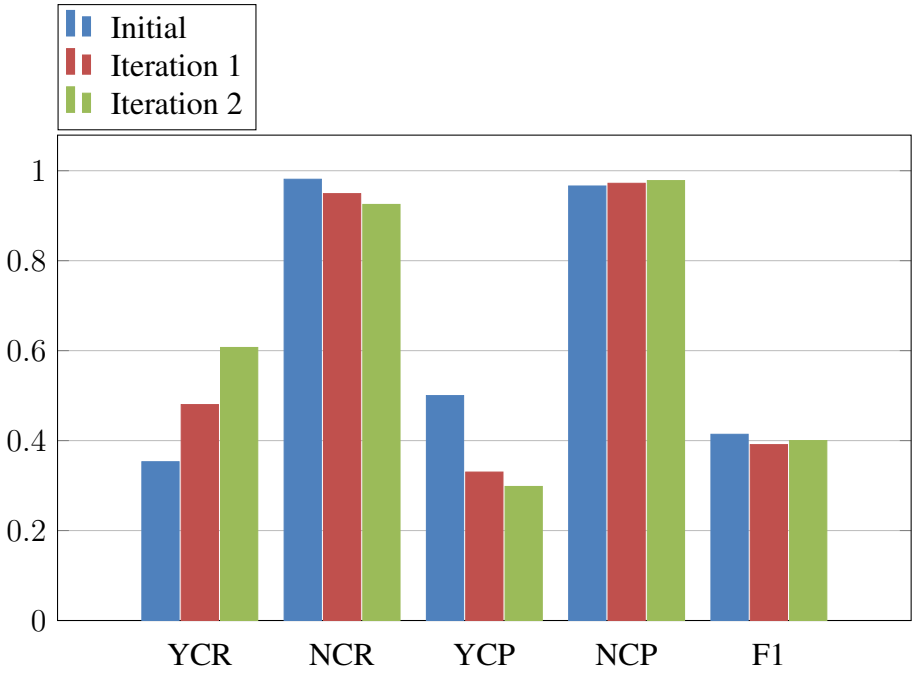

Figure 4.5: Random walk iterations with hub weight 0.01 and authority weight 0.01

4.1.2 Only User Initial Scoring Results

In Figure 4.5 YCR showed increase as iteration count increased. Iteration 2 YCR showed 80% increase in Iteration 2 experiment than Initial experiment. On the other hand YCP 50%. NCR, NCP and F1 scores changed slightly.

In Figure 4.6, similar to in Figure 4.5, YCR increased 78%, YCP decreased 62% and NCP did not change much between Initial and Iteration 2 experiments. F1 scores of Iteration 2 experiment showed 13% decrease than Initial experiment.

In Figure 4.7, Iteration 2 and 1 experiments gave close results where YCR is 89 better than Initial experiment, YCP decreased by 56%, NCR decreased by %8, NCP slightly changed and F1 score decreased by 20%.

In Figure 4.8, the best YCR is seen in the Iteration 1 experiment. Its YCR result showed 95% improvement while Iteration 2 YCR showed only 60% improvement than Initial experiment. On the other hand YCP scores decreased by 53% resulting worse F1 scores in the experiments. Iteration 2 experiment has better F1 score than Iteration 1 and its results showed 25% decrease when we compared it with Initial

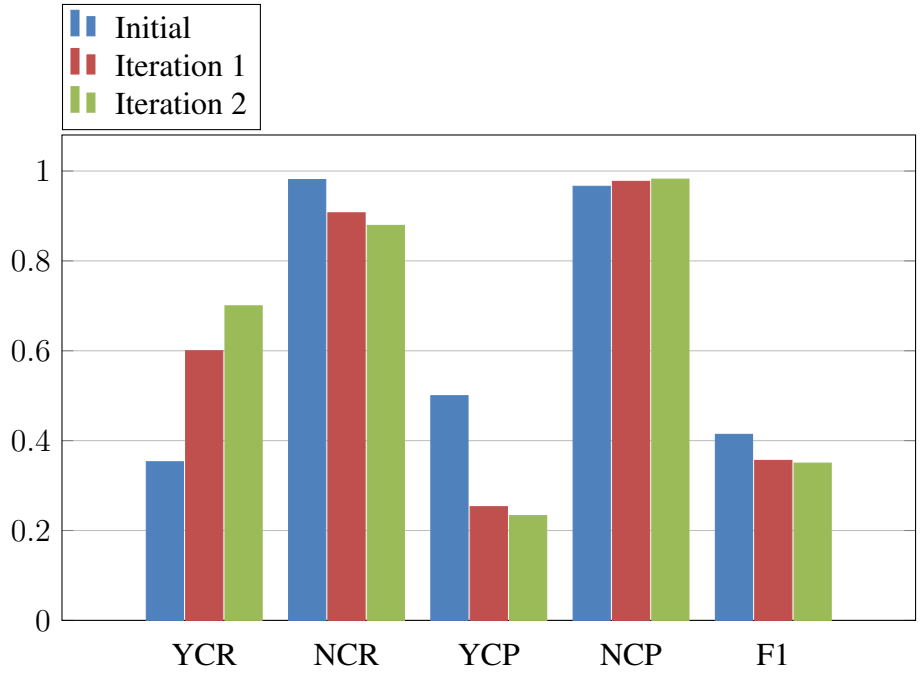

Figure 4.6: Random walk iterations with hub weight 0.04 and authority weight 0.04

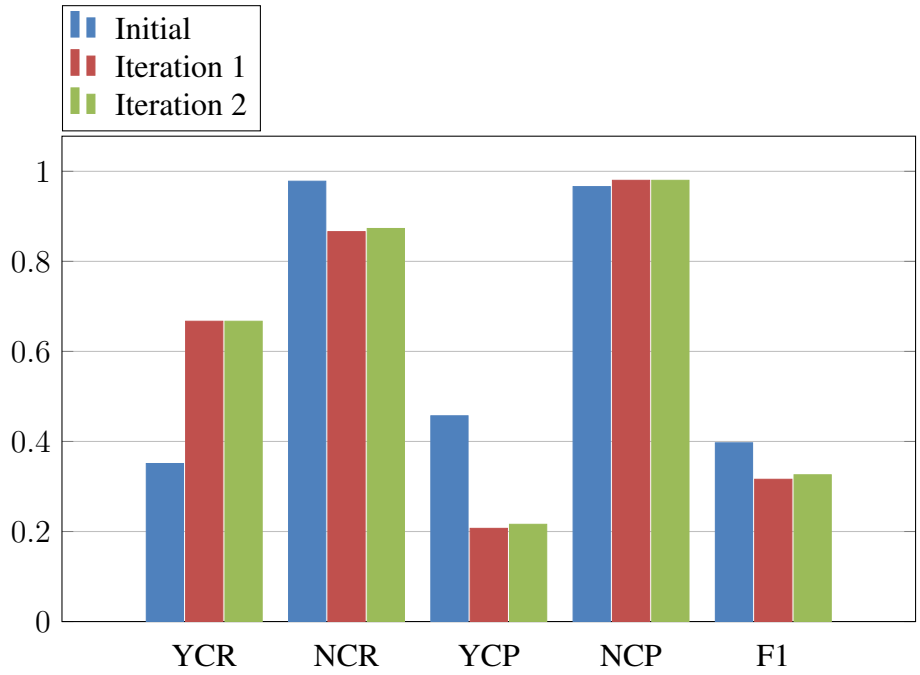

Figure 4.7: Random walk iterations with hub weight 0.07 and authority weight 0.07

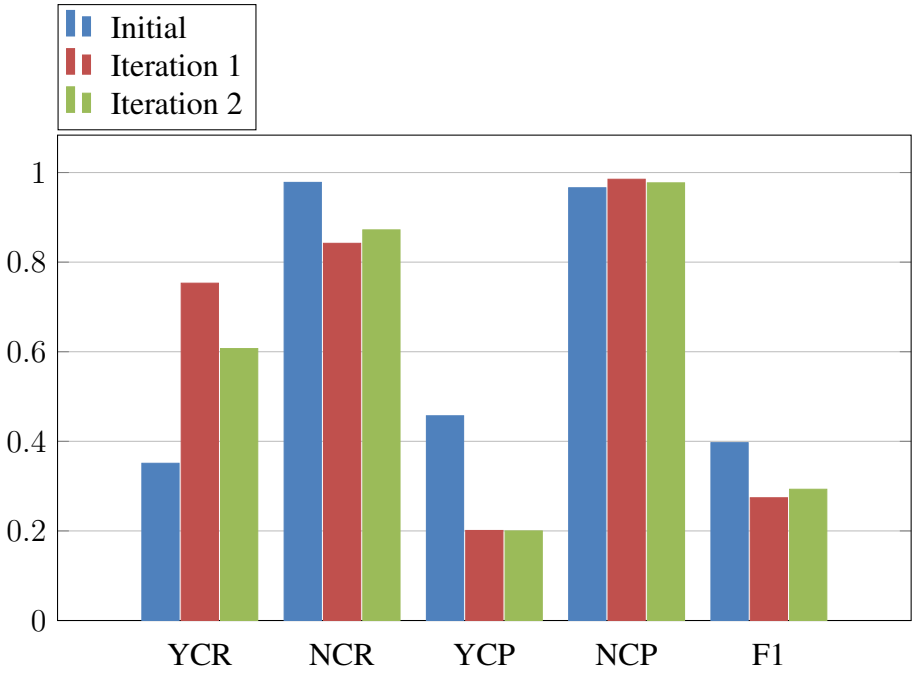

Figure 4.8: Random walk iterations with hub weight 0.1 and authority weight 0.1

experiment.

When we compared the results of iteration experiments with the initial results, we did not see much improvement. Best F1 score of the iteration experiments appeared to be with 0.01 weights. Giving positive and negative initial scores to users in our graph structure lead to better initial experiment performance which is deteriorated as random walk iterations applied on it. Making experiments with changing weights showed that giving initial scores to user nodes while not scoring the tweet nodes is not a successful attempt and its results appeared to be independent of weights.

4.1.3 User and Tweet Hybrid Initial Scoring Results

Both user and tweet nodes are initially scored in this experiment and score transfer weight 0.01 is used. We selected hub and authority weights 0.01 since it is appeared to be the best weight in the experiments explained in Section 4.1.1 and Section 4.1.2. This experiment is a hybrid of aforementioned section's experiments.

In Figure 4.9, we see that YCR increased as iteration count increase but YCP de-

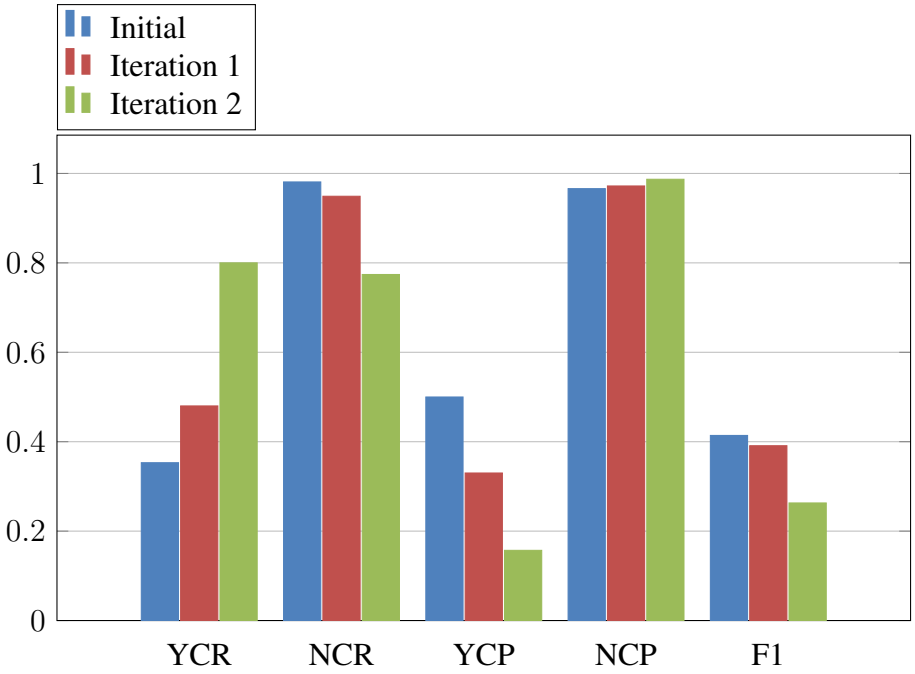

Figure 4.9: Random walk iterations with hub weight 0.01 and authority weight 0.01 with user and tweet hybrid initial scoring

creased. Similarly NCR results decreased while NCP slightly changed. Iteration 1 F1 score appeared to be better than Iteration 2 F1 score and it is 5% worse than Initial experiment. Hybridizing with user initial scoring resulted in worse F1 score performance when we compare with Figure 4.1.

4.1.4 Dictionary Based Analysis Results

For first dimension of credibility we made 6 more experiments which are:

- E2, only considering word existence in slang word dictionary of tweet text
- E3, considering both word existence in slang word dictionary of tweet text and first phase result label of the tweet
- E4, considering intersection of the tweets which has word existence in slang word dictionary and positive first phase result label
- E5, considering the tweets whose negative sentiment score <= -3

- E6, considering both the tweets whose negative sentiment score <= -3 and the tweets having word existence in slang word dictionary

- E7, considering the intersection of the tweets whose negative sentiment score <= -3 and the tweets having word existence in slang word dictionary

E1 is the experiment results of random walk iterations with hub weight 0.01 and authority weight 0.01 which has the best F1 scores among other graph based improvement experiments for the first dimension. We compared the other results with E1.

Table4.1: YCR,NCR,YCP,NCP and F1 score results of experiments

Experiment:	E1	E2	E3	E4	E5	E6	E7
Yes Class Recall:	0,413	0,753	0,787	0,36	0,373	0,353	0,34
No Class Recall:	0,966	0,954	0,934	0,995	0,942	0,989	0,993
Yes Class Precision:	0,392	0,461	0,384	0,794	0,252	0,631	0,729
No Class Precision:	0,969	0,987	0,988	0,967	0,966	0,967	0,966
F1 Score:	0,403	0,571	0,516	0,495	0,301	0,453	0,464

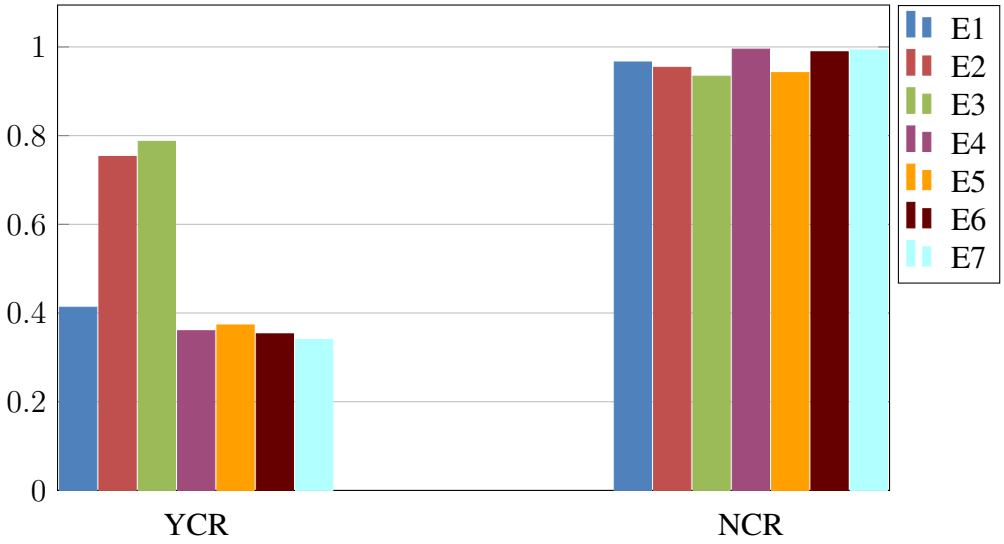

Figure 4.10: YCR and NCR Results of experiments for the first dimension

As it can be seen in the Table 4.10 best YCR is obtained in E3 and best NCR is obtained in E4.

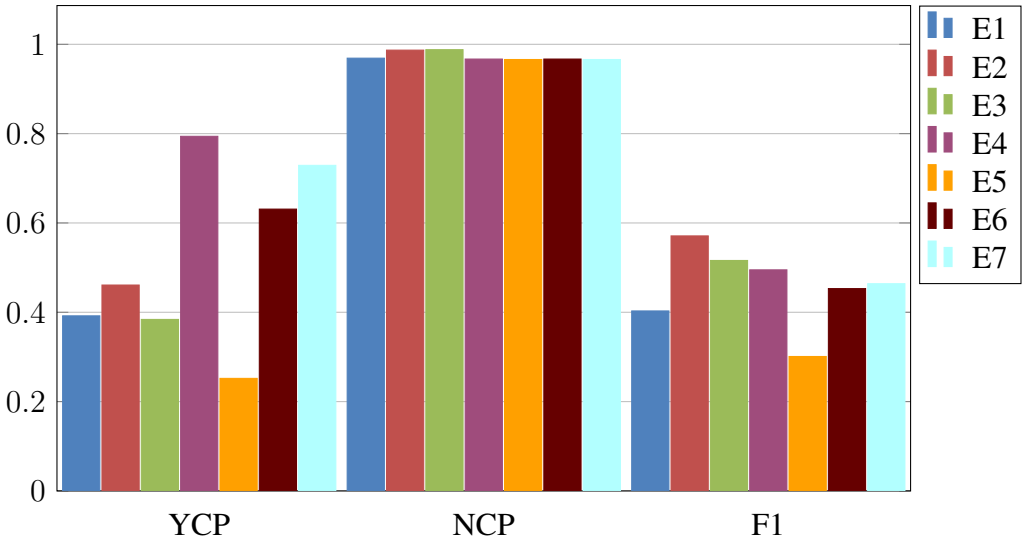

Figure 4.11: YCP, NCP and F1 score results of experiments for the first dimension

Figure 4.11 shows that best F1 score is obtained in E2, best NCP is obtained in E3 and best YCP is obtained in E4.

4.2 Experimental Analysis For Dimension 2 - Spam Tweets

The second dimension is about filtering spam tweets. To check this, in the user study, volunteers were asked the following question: ”Is the tweet written for distracting, unrelated, advertising or out of program scope purposes?” (In Turkish ”Dikkat Dağıtıcı, Alakasız, Reklam İçerikli veya Program Dışı Bir Amaçla mı Yazılmış?”). The ground truth for this data set includes 393 positive and 2607 negative tweets.

4.2.1 Only Tweet Initial Scoring Results

In Figure 4.12, in Iteration 2 experiment YCR increased to 16% and YCP decreased by 27%. F1 score of Iteration 1 is better than Iteration 2 experiment but nonetheless

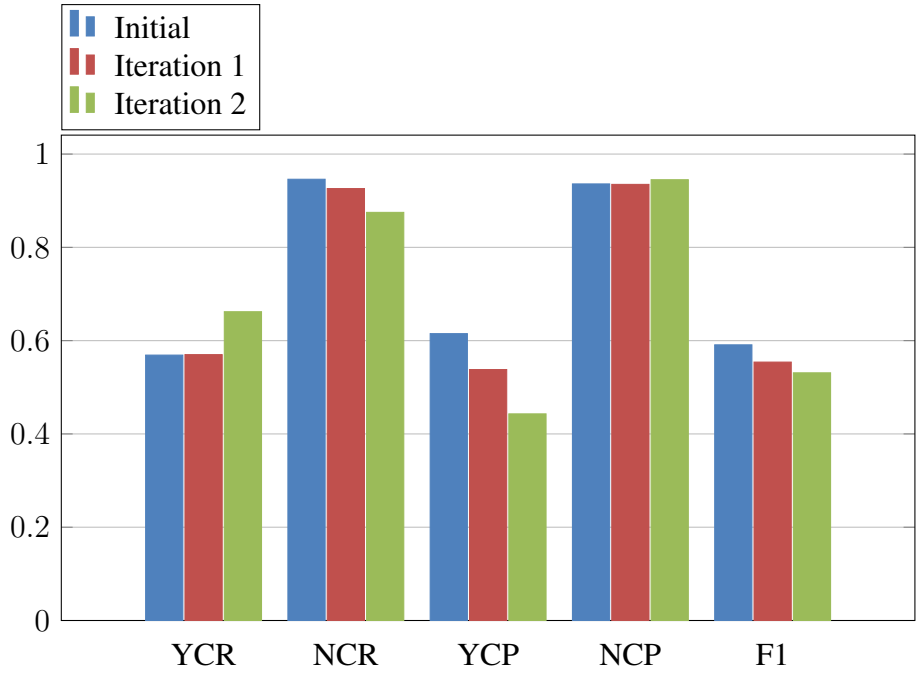

Figure 4.12: Random walk iterations with hub weight 0.01 and authority weight 0.01

it is 6% worse than Initial experiment.

In Figure 4.13 we observed YCR scores increased as iteration counts increase however NCR, YCP and F1 scores decreased. Initial experiment F1 score appeared to be 13% better than Iteration 1 experiment.

In Figure 4.14 we observed similar results patterns with Figure4.13.

In Figure 4.15 we observed similar results patterns with Figure4.13.

In iteration based experiments for the second dimension, the best F1 score is observed with weight 0.01 iterations. However all of the experiments with weights 0.01 to 0.1 resulted worse F1 scores while resulting similar bar chart patterns. We observed that iterations did not improve the classification performance of feature based phase. This showed that internal structure of our node-user graph lacked the ability to separate spam tweets from credible ones.

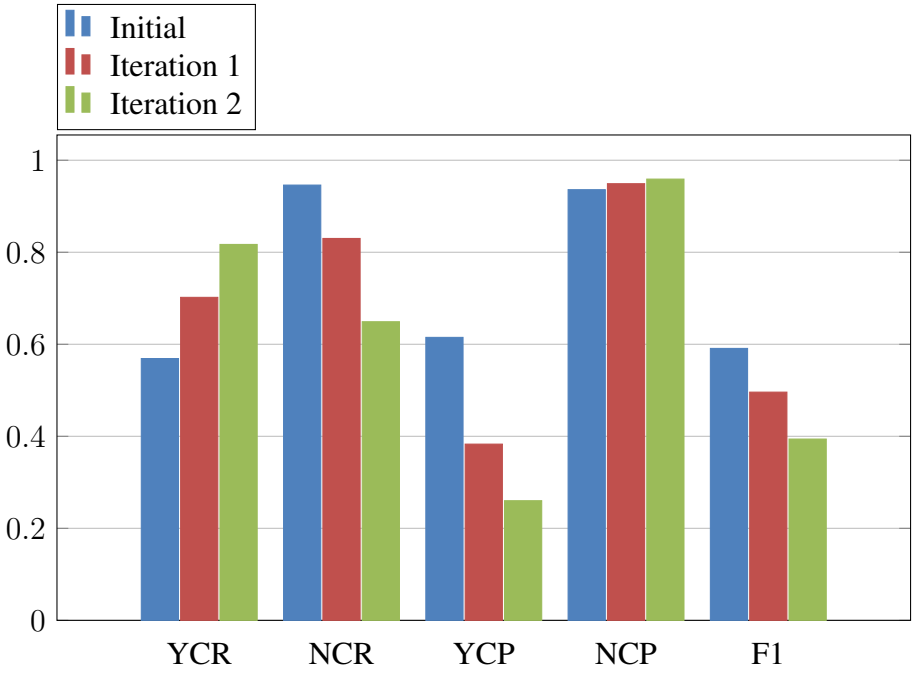

Figure 4.13: Random walk iterations with hub weight 0.04 and authority weight 0.04

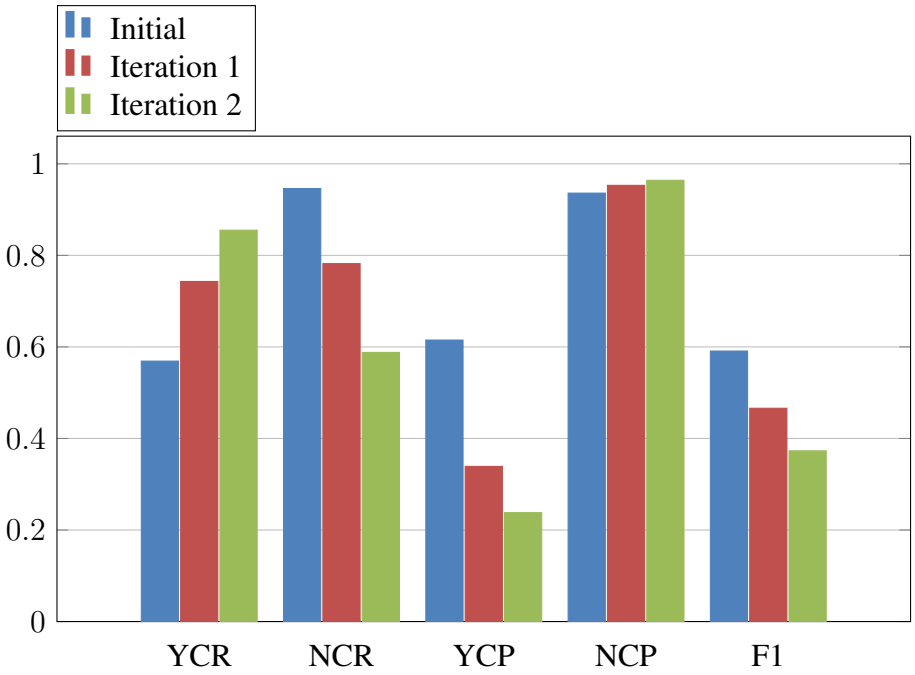

Figure 4.14: Random walk iterations with hub weight 0.07 and authority weight 0.07

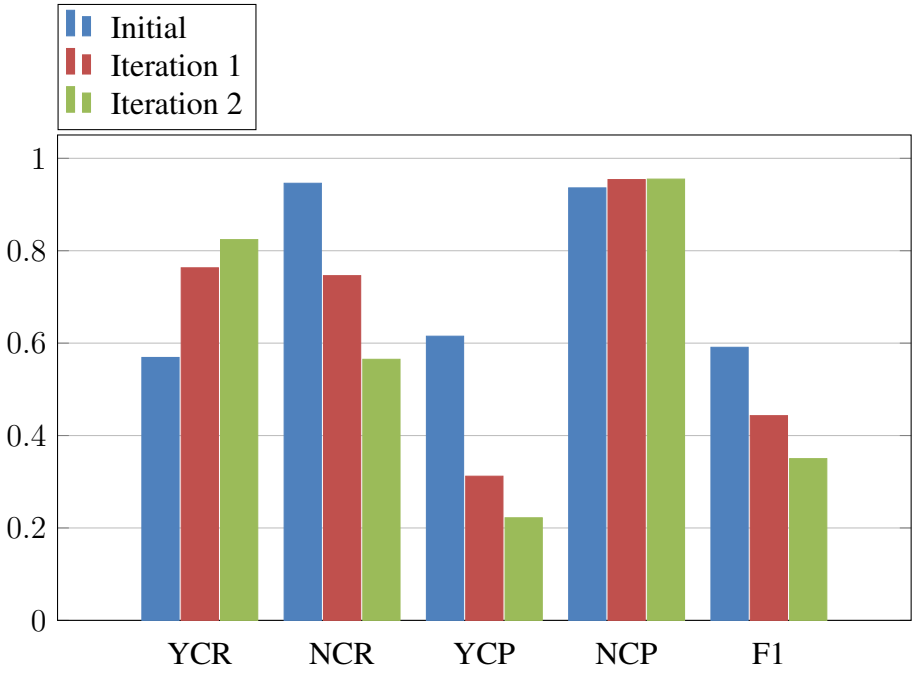

Figure 4.15: Random walk iterations with hub weight 0.1 and authority weight 0.1

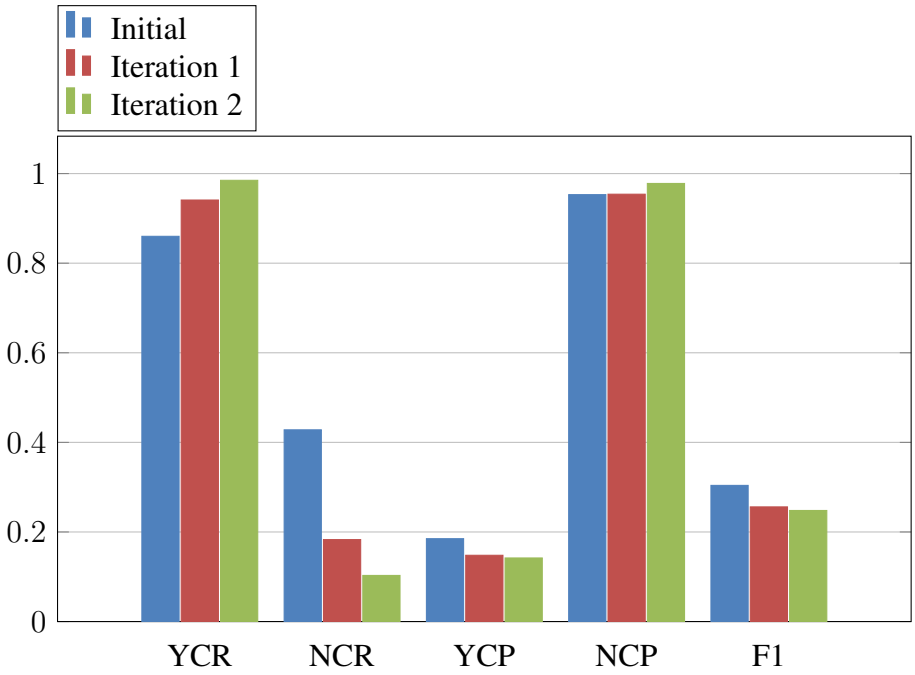

Figure 4.16: Random walk iterations with hub weight 0.01 and authority weight 0.01

4.2.2 Only User Initial Scoring Results

In Figure 4.16, YCR improved by 13% in Iteration 2 experiment with respect to Initial experiment however NCR decreased by 76%. F1 score of Iteration 2 experiment is 18% worse than Initial experiment.

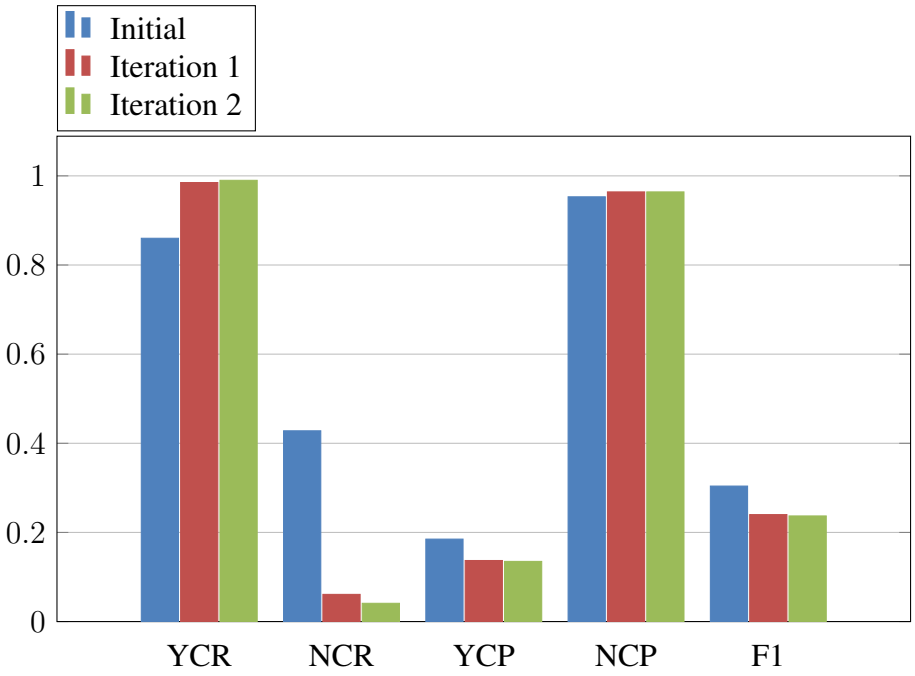

Figure 4.17: Random walk iterations with hub weight 0.04 and authority weight 0.04

In Figure 4.17 Iteration 1 and 2 showed very close results. F1 score difference between them is less than 1%. However Initial experiment F1 score did not improve but decreased 20%.

In Figure 4.18 we observed similar results patterns with Figure4.17.

In Figure 4.19 we observed similar results patterns with Figure4.18.

Similar to Section 4.2.1 experiment results showed that the internal graph structure lacked the ability to separate spam tweets from credible ones. Moreover giving initial scores to users and leaving tweets appeared to be less effective than Section 4.2.1. This showed that users can write both kind of tweets which can not be classified only depending on the writer of them.

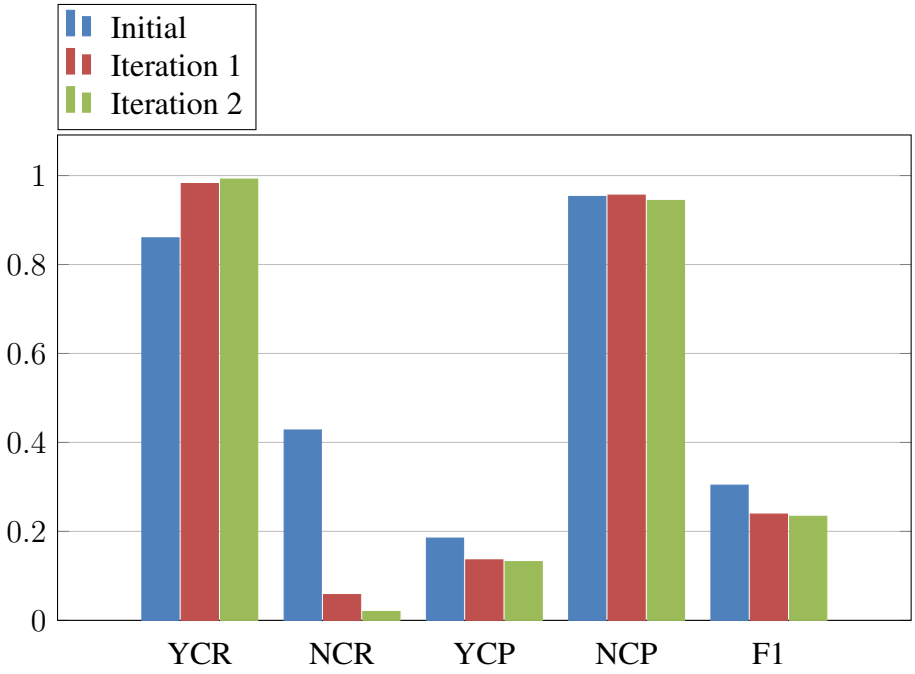

Figure 4.18: Random walk iterations with hub weight 0.07 and authority weight 0.07

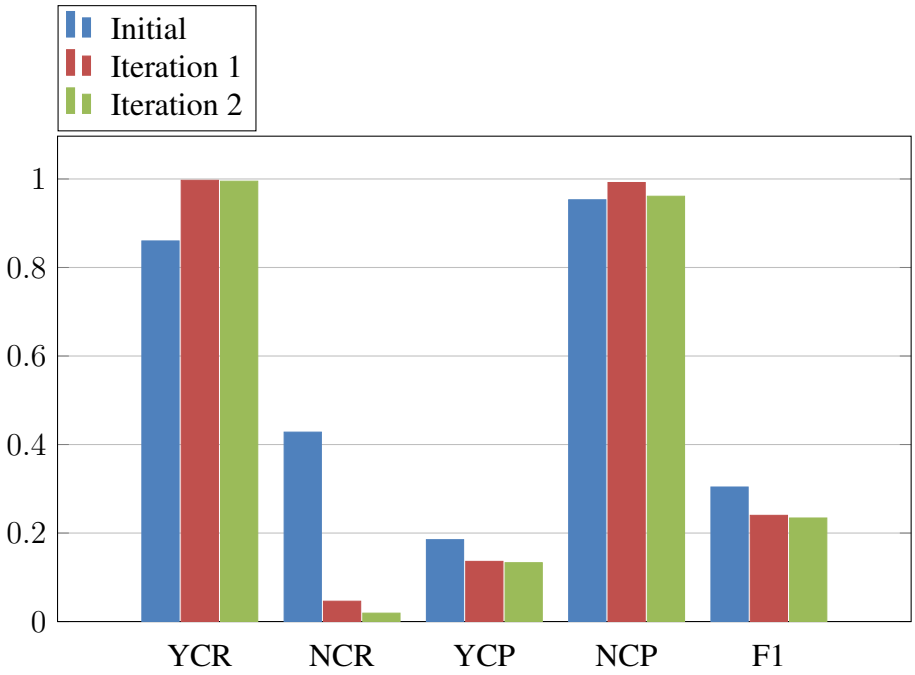

Figure 4.19: Random walk iterations with hub weight 0.1 and authority weight 0.1

4.2.3 User and Tweet Hybrid Initial Scoring Results

The best results of random walk iterations for tweet initial scoring are obtained with weights 0.01 so we made experiments with that value with hybrid initial scoring for the second question.

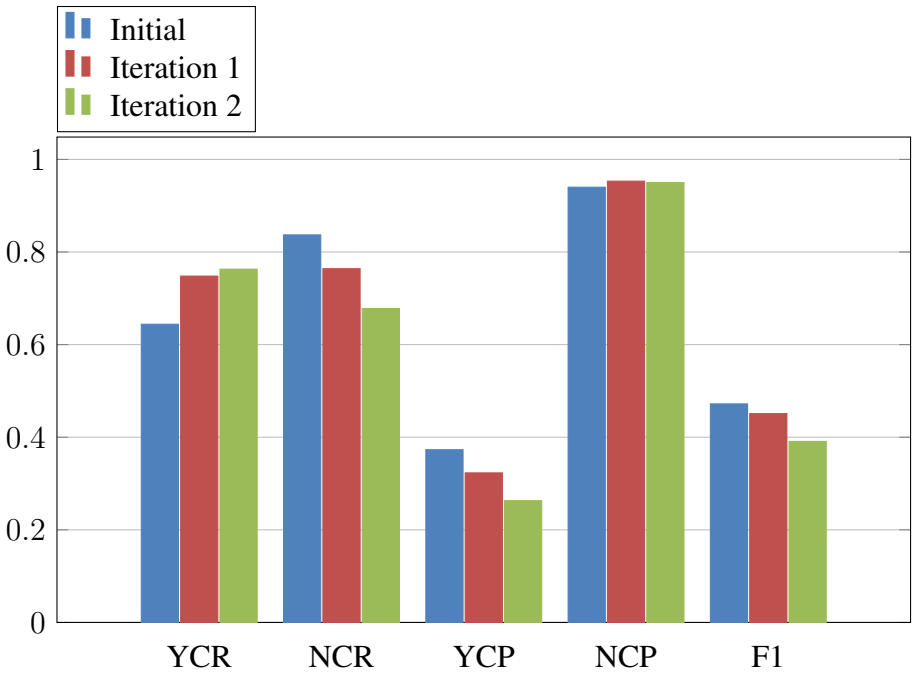

Figure 4.20: Random walk iterations with hub weight 0.01 and authority weight 0.01 with user and tweet hybrid initial scoring

In Figure 4.20, in Iteration 2 experiment YCR increased to 18%, NCR descreased by 19%, YCP decreased by 17% and F1 score decreased by 21% with respect to Initial experiment results.

Hybrid initial scoring increased the F1 scores with respect to only user initial scoring however it did not exceed the results of first phase for the second dimension.

4.3 Experimental Analysis For Dimension 3 - Newsworthiness

The third dimension is about the news-worthiness. To check this, in the user study, volunteers were asked the following question: ”Is the content interesting, important or news-worthy?” (In Turkish ”İçerik İlginç, Dikkate Değer veya Haber Değeri Taşıyor mu?”). The ground truth for this data set includes 2068 positive and 932 negative tweets.

4.3.1 Only Tweet Initial Scoring Results

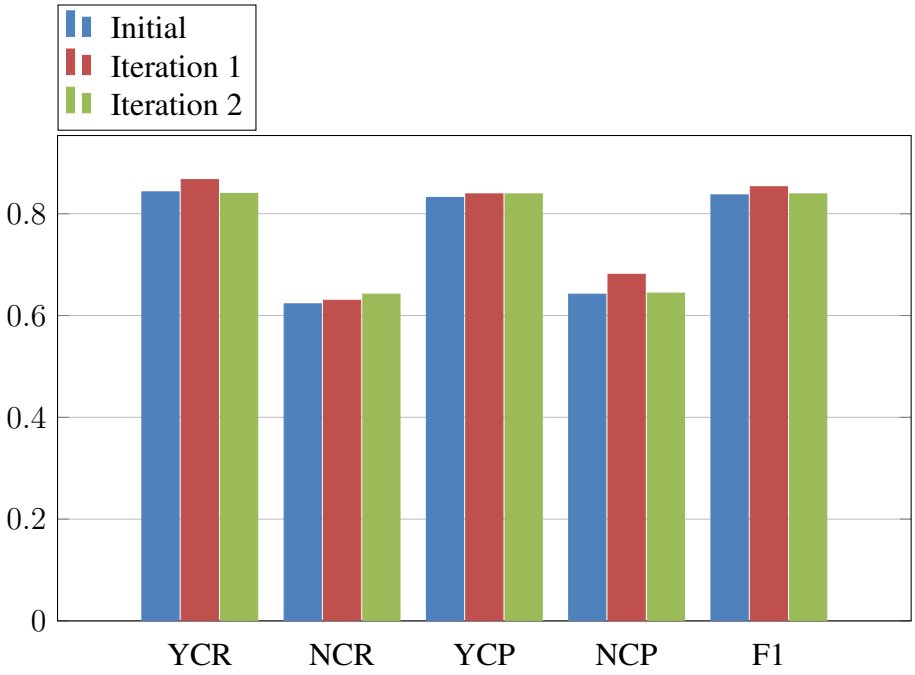

Figure 4.21: Random walk iterations with hub weight 0.01 and authority weight 0.01

In Figure 4.21, YCR, NCR, YCP and NCP scores slightly changed. Best F1 score is obtained in the Iteration experiment which barely improved initial performance by 1%.

In Figure 4.22, YCR decreased by 17% and NCP decreased by 18% while NCR increased to 12% and YCP increased to 3% in Iteration 2 experiment with respect to Initial experiment. In addition we observed that F1 score decreased by 8%.

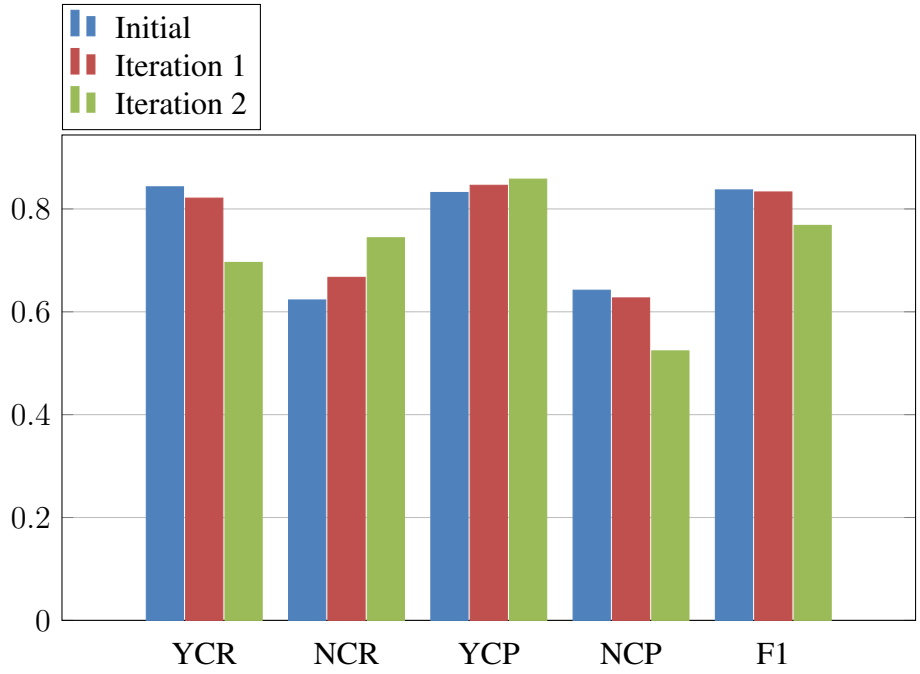

Figure 4.22: Random walk iterations with hub weight 0.04 and authority weight 0.04

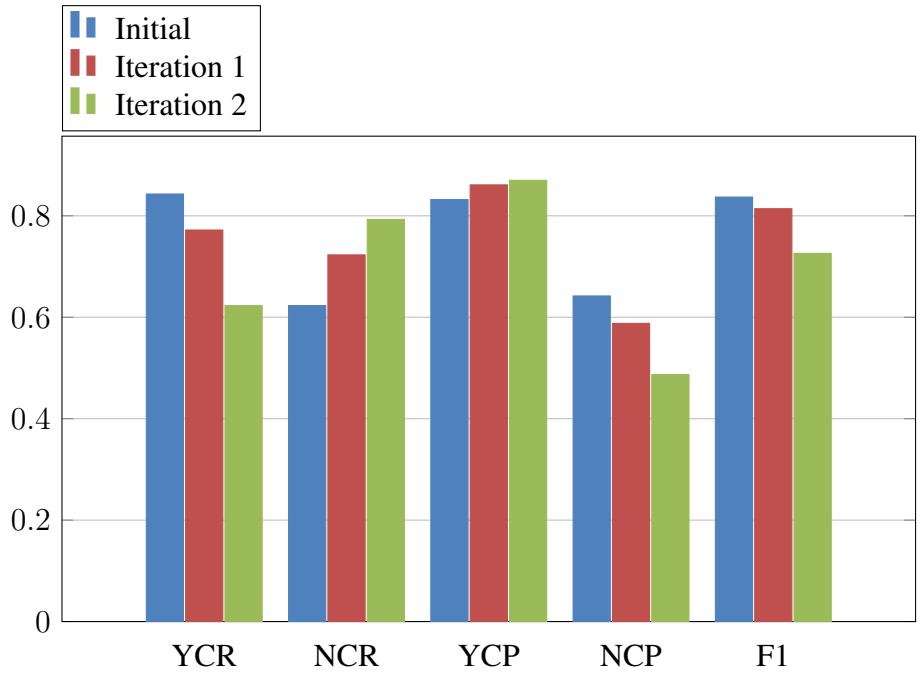

Figure 4.23: Random walk iterations with hub weight 0.07 and authority weight 0.07

In Figure 4.23, YCR decreased by 26%, NCP decreased by 24% and F1 decreased by 13% in Iteration 2 experiment with respect to Initial experiment. On the other hand NCR increased to 27% and YCP increased to 4%.

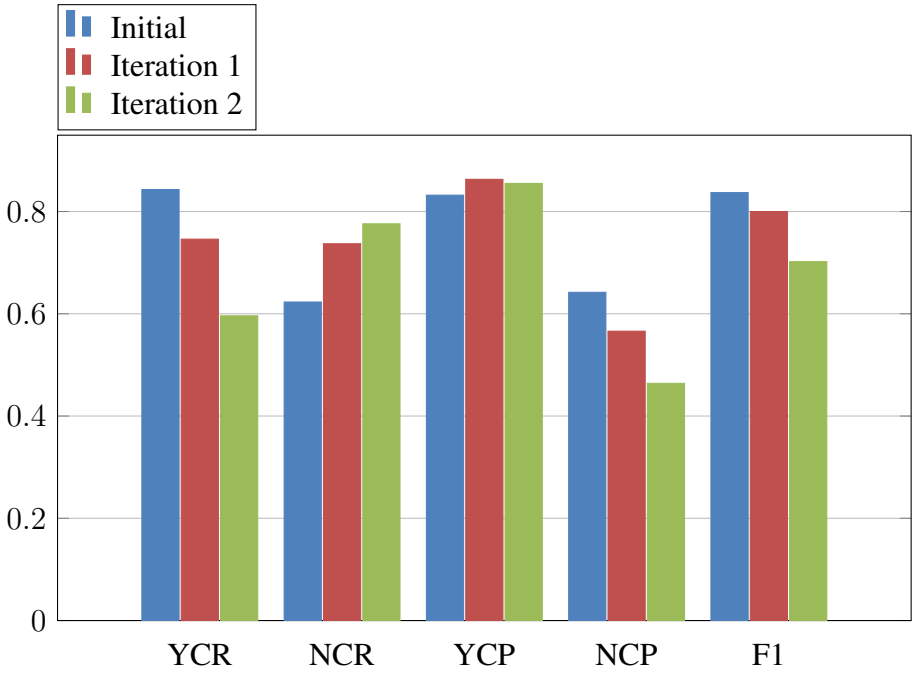

Figure 4.24: Random walk iterations with hub weight 0.1 and authority weight 0.1

In Figure 4.24 we observed similar results patterns with Figure 4.23.

During those experiment we obtained the best F1 scores with the 0.01 hub and authority weight experiment for the third dimension. Increasing weights did not help improvement performance due to internal structure of our graph showing that positive tweets are linked with negative tweets as well. As larger scores passed between nodes, we observed that newsworthy tweets are concealed by the negative ones in the graph.

4.3.2 Only User Initial Scoring Results

In Figure 4.25, we observed that Iteration 1 and 2 experiment results appeared very close. YCR results approached to 1 and NCR results approached 0 since true nega-

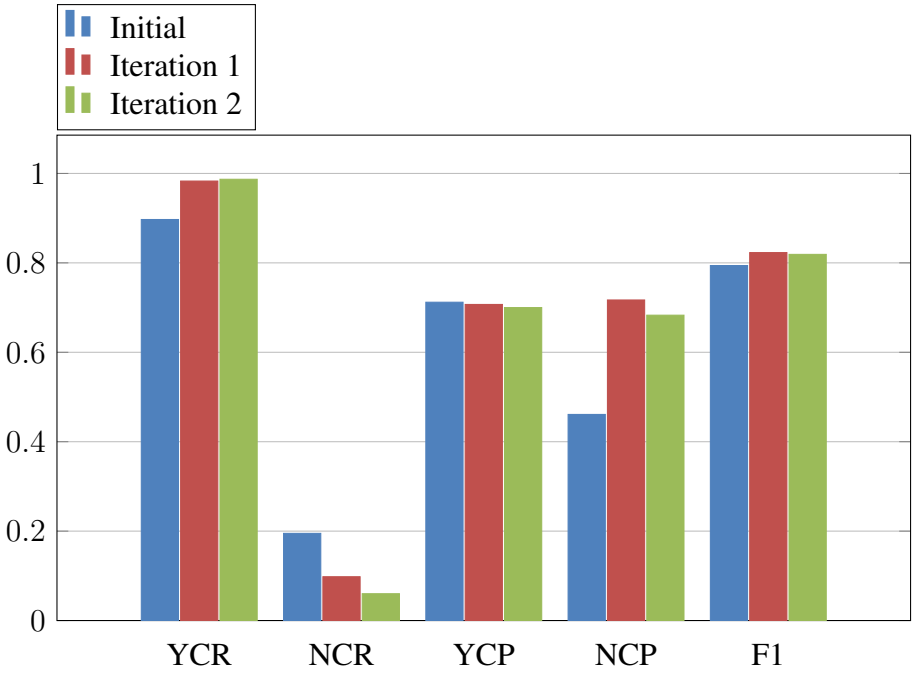

Figure 4.25: Random walk iterations with hub weight 0.01 and authority weight 0.01

tive counts are decreased and false positive counts increased as it can be seen in the Table C.5.

In Figure 4.26 we observed similar results patterns with Figure 4.25.

In Figure 4.27 we observed similar results patterns with Figure 4.26.

In Figure 4.28 we observed similar results patterns with Figure 4.27.

We observed that scoring users and giving 0 initial scores to tweets nodes showed an increase in F1 score as iterations increased unlike scoring only tweet nodes and giving 0 initial scores to user nodes approach. The best F1 scores are observed with weight 0.1 during these experiments for this question. However as it can be seen from the Figure 4.21, tweet initial scoring with 0.01 hub and authority weight yielded better results. As the same thing is observed in the previous dimension experiments, this is mainly caused from the internal connection structure of our graph.

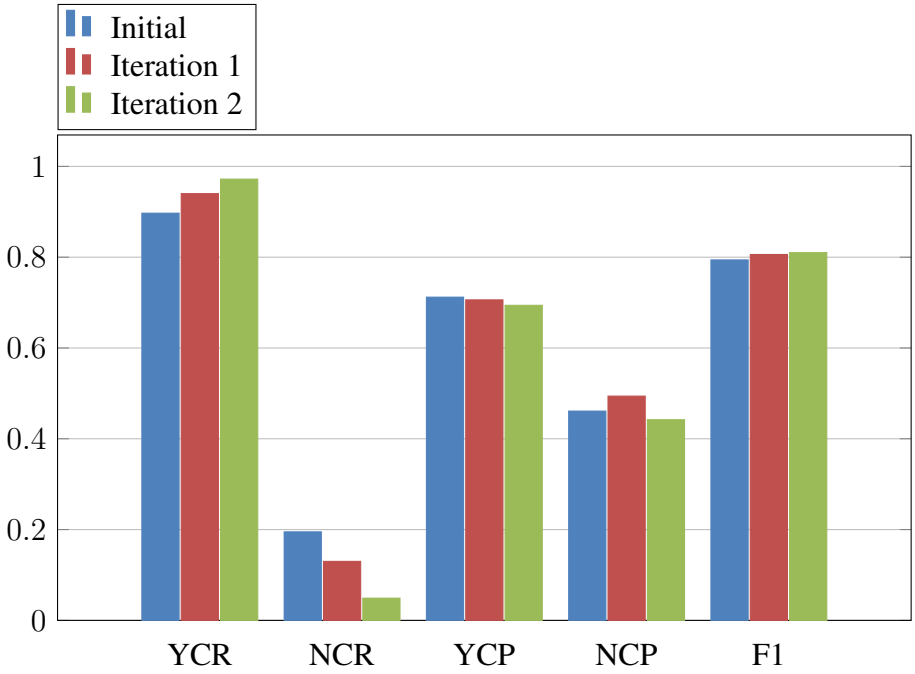

Figure 4.26: Random walk iterations with hub weight 0.04 and authority weight 0.04

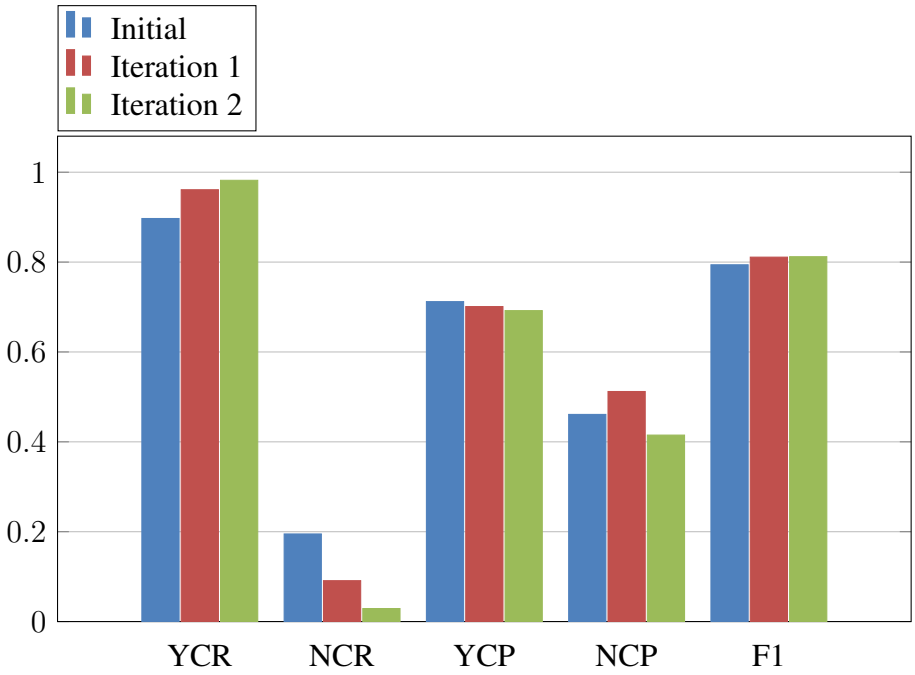

Figure 4.27: Random walk iterations with hub weight 0.07 and authority weight 0.07

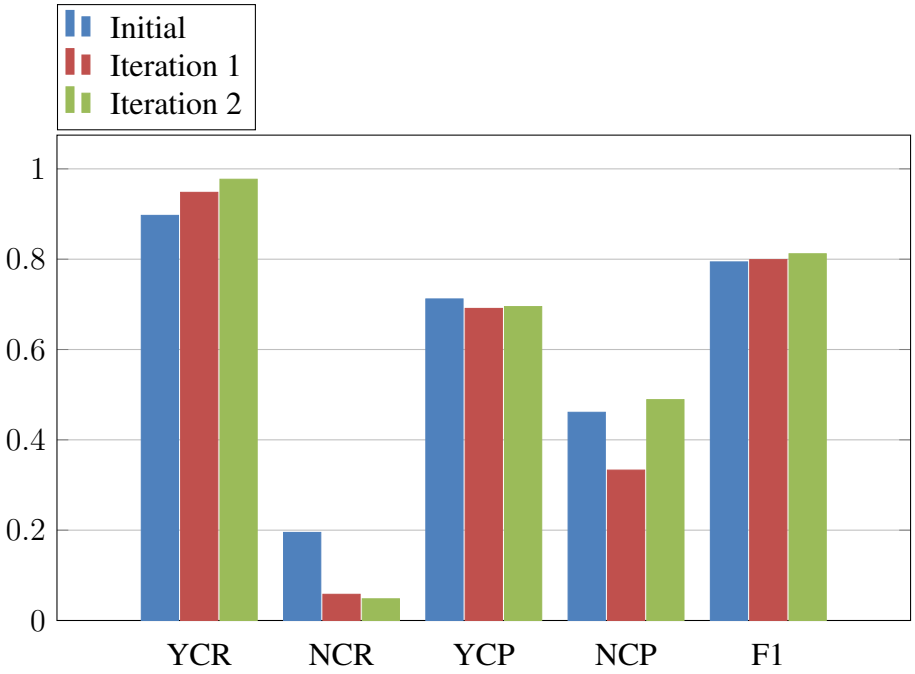

Figure 4.28: Random walk iterations with hub weight 0.1 and authority weight 0.1

4.3.3 User and Tweet Hybrid Initial Scoring Results

The best results of random walk iterations for tweet initial scoring are seen with weights 0.01 so we made experiments with that value with hybrid initial scoring for the second question.

In Figure 4.29, YCR increased to 13% and NCP increased to 18% in Iteration 2 experiment with respect to Initial experiment. On the other hand, NCR decreased by 52%, YCP decreased by 10% and F1 score did not change more than 1%.

The experiment with third question showed that hybrid initial scoring F1 scores did not change as iteration count increase. Moreover its F1 scores appeared to be better than all of the other experiments with third question experiments other than iteration 1 with only tweet initial scoring experiment. Only here we saw better performance and improvement of results of first phase feature based classification.

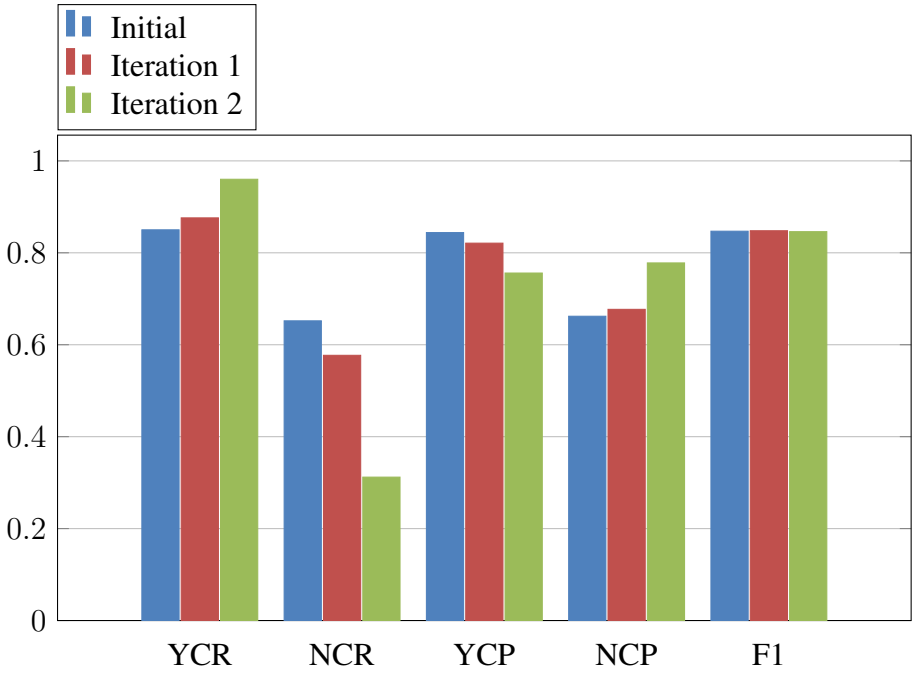

Figure 4.29: Random walk iterations with hub weight 0.01 and authority weight 0.01 with user and tweet hybrid initial scoring

4.4 Overall Credibility Decision

In this section we present the credible tweets according to our experiment results of both three dimensions. In Table 4.2 programs, total number of tweets written for each program, number of credible tweets and their averages are shown. As explained in our credibility definition in Section 3.1 we labelled a tweet as credible only if it belongs to no class of first dimension and second dimension and yes class of third dimension. We based our final credibility decision upon the results of experiments having best F1 scores. We produced the final credibility results from E2 experiment for first dimension, 0.01 weight only tweet initial scoring experiment for second dimension and 0.01 weight only tweet initial scoring experiment for the third dimension in this part.

Table4.2: Credibility of tweets per programs

No	Program Name	Total Tweets	No of Predicted to be Credible Tweets	No of Credible Tweets According to User Annotation
1	Deşifre	370	291	250
2	5N 1K	272	178	102
3	Son Söz	218	119	52
4	Tarafsız Bölge	213	163	83
5	%100 Siyaset	176	109	90
6	Oteki Gündem	174	119	118
7	Karşıt Görüş	165	105	117
8	Yaz Boz	154	97	111
9	Tarihin Arka Odası	147	115	58
10	Türkiyenin Nabzı	140	112	76
11	Dinamit	136	85	38
12	Kadraj	126	98	64
13	Memleket Meselesi	100	61	65
14	Ne Oluyor	76	48	43
15	Bıçak Sırtı	76	56	37
16	Ceviz Kabuğu	74	48	35
17	Yakın Plan	72	35	24
18	Çıkış Yolu	72	47	46
19	Birlikte Bakalım	69	50	35

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this thesis, we studied tweet credibility problem. We defined credibility as being free from offensive words, being dedicated to the purpose of the discussion and being newsworthy. We collected tweet and user data from Twitter. Those tweets are written for Turkish news and discussion programs which have contents ranging from social, economic and cultural topics to political discussions. To justify our proposed method performance we collected ground truth data from human volunteers through our website and get the tweets voted by at least three voters.

We proposed a method to investigate credibility based on the relation between tweet-tweet, user-user and tweet-user network in the light of feature based machine learning methods. In the supervised learning phase of our method we predicted yes/no class labels of our data set by J48 decision tree classifier. Those initial prediction's recall and precision results are improved in the second phase of our proposed method. By creating a graph from user and tweet nodes and linking them according to tweet-tweet word similarity, user-user friend/follower relation and tweet-user owner connection we searched whether credible tweets are linked with each other or not. We aimed to separate positive and negative classes by applying hub/authority score transfers in the graph. Moreover we supported our method by deploying slang word dictionary checking algorithms. Even though we developed a method based on Turkish language, our study can be generalized for other languages by changing language parser and word separator components.

Any tweet is considered as credible finally if it is free from slang words, dedicated to the program content and important/news-worthy. Those three dimensions are examined separately in supervised learning and graph based improvement phases and labels are produced independently. We made experiments and discussed those results for those three dimensions in Chapter 4.

In the graph based improvement phase, we made experiments by giving initial scores to only tweets, to only users and both users and tweets. We looked at YCR, NCR, YCP, NCP and F1 score results among those experiments in Chapter 4.

Our experiment results showed that deploying a slang word dictionary increases the precision, recall and F1 score results. We improved 91% the YCR, 3% NCR, 102% YCP, 2% NCP and 41% F1 score in dictionary utilized experiments in the first dimension experiments.

In the second dimension experiments we achieved to increase YCR 18% and NCP 1% however we obtained poorer F1 scores.

On the other hand, in third dimension experiments we succeeded to increase F1 score 1%. In those experiments we observed that NCR increased 12% and YCP increased 4% as well.

5.2 Future work

This study can be improved further by increasing iteration numbers of experiments and changing hub/authority weights. In our study we made experiments with 0.01, 0.04, 0.07 and 0.1 weight experiments for both hub and authority weight pairs. Those rational scores set can be enriched. Beyond this, we can make experiments with different hub and authority scoring combinations.

Moreover data size can be increased as well. Increasing data size will provide better results and would enable us to propose more robust solution to credibility analysis problem. More users and more tweets would provide a more realistic work set but this data set need more human voters to obtain ground truth values. For this purpose, we need to hire volunteers and make questionnaire studies with them.

In addition, we can generalize our proposed method for other languages. This study is only focused on Turkish language however by changing language parser and word separator, we can develop this study for other languages as well.

REFERENCES

[1] M.-A. Abbasi and H. Liu. Measuring user credibility in social media. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 441–448. Springer, 2013.

[2] 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*, page 2. ACM, 2013.

[3] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 564–575. VLDB Endowment, 2004.

[4] I.-A. Bara, C. J. Fung, and T. Dinh. Enhancing twitter spam accounts discovery using cross-account pattern mining. In *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, pages 491–496. IEEE, 2015.

[5] P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.

[6] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *COMPUTER NETWORKS AND ISDN SYSTEMS*, pages 3825–3833. Elsevier Science Publishers B. V., 1998.

[8] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[9] E. M. Clark, J. R. Williams, R. A. Galbraith, C. M. Danforth, P. S. Dodds, and C. A. Jones. Sifting robotic from organic text: A natural language approach for detecting automation on twitter. *arXiv preprint arXiv:1505.04342*, 2015.

[10] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *arXiv preprint arXiv:1407.5225*, 2014.

[11] B. Fogg and H. Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87. ACM, 1999.

[12] B. J. Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pages 722–723. ACM, 2003.

[13] M. Forelle, P. Howard, A. Monroy-Hernández, and S. Savage. Political bots and the manipulation of public opinion in venezuela. *arXiv preprint arXiv:1507.07109*, 2015.

[14] A. Gün and P. Karagöz. A hybrid approach for credibility detection in twitter. In *Hybrid Artificial Intelligence Systems*, pages 515–526. Springer, 2014.

[15] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama. Assessment of tweet credibility with lda features. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 953–958. International World Wide Web Conferences Steering Committee, 2015.

[16] B. Kang, T. Höllerer, and J. O'Donovan. Believe it or not? analyzing information credibility in microblogs. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 611–616. ACM, 2015.

[17] 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*, pages 179–188. ACM, 2012.

[18] A. M. Kaplan and M. Haenlein. The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2):105–113, 2011.

[19] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[20] S. Kumar, F. Morstatter, R. Zafarani, and H. Liu. Whom should i follow?: identifying relevant users during crises. In *Proceedings of the 24th ACM conference on Hypertext and social media*, pages 139–147. ACM, 2013.

[21] J. Martinez-Romo and L. Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000, 2013.

[22] 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*, pages 1155–1158. ACM, 2010.

[23] M. Mccord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Autonomic and trusted computing*, pages 175–186. Springer, 2011.

[24] J. O'Donovan, B. Kang, G. Meyer, T. Hollerer, and S. Adalii. Credibility in context: An analysis of feature distributions in twitter. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 293–301. IEEE, 2012.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[26] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM, 2011.

[27] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, pages 265–272, 2011.

[28] 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*, pages 851–860. ACM, 2010.

[29] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[30] A. H. Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10. IEEE, 2010.

[31] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In *Web Information Systems Engineering–WISE 2010*, pages 240–253. Springer, 2010.

[32] C. Yang, R. C. Harkreader, and G. Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Recent Advances in Intrusion Detection*, pages 318–337. Springer, 2011.

APPENDIX A

EXPERIMENTAL ANALYSIS FOR DIMENSION 1 -SLANG LANGUAGE

TableA.1: Only tweet initial scoring random walk iterations with hub weight 0,01 and authority weight 0,01

Experiment:	Iteration=0	Iteration=1	Iteration=2
False Positive:	62	96	196
True Negative:	2788	2754	2654
True Positive:	53	62	72
False Negative:	97	88	78
Yes Class Recall:	0,351	0,413	0,480
No Class Recall:	0,978	0,966	0,931
Yes Class Precision:	0,457	0,392	0,269
No Class Precision:	0,966	0,969	0,971
F1 Score:	0,397	0,403	0,344
Accuracy:	0,947	0,938	0,909
Specificity:	0,978	0,966	0,931
Sensitivity:	0,353	0,413	0,480

TableA.2: Only tweet initial scoring random walk iterations with hub weight 0,04 and authority weight 0,04

Experiment:	Iteration=0	Iteration=1	Iteration=2
False Positive:	64	258	360
True Negative:	2786	2592	2490
True Positive:	53	96	105
False Negative:	97	54	45
Yes Class Recall:	0,353	0,640	0,700
No Class Recall:	0,978	0,909	0,874
Yes Class Precision:	0,453	0,271	0,226
No Class Precision:	0,966	0,980	0,982
F1 Score:	0,397	0,381	0,341
Accuracy:	0,946	0,896	0,865
Specificity:	0,977	0,909	0,873
Sensitivity:	0,353	0,640	0,700

TableA.3: Only tweet initial scoring random walk iterations with hub weight 0,07 and authority weight 0,07

Experiment:	Iteration=0	Iteration=1	Iteration=2
False Positive:	64	387	372
True Negative:	2786	2463	2478
True Positive:	53	99	97
False Negative:	97	51	53
Yes Class Recall:	0,353	0,660	0,647
No Class Recall:	0,978	0,864	0,869
Yes Class Precision:	0,453	0,204	0,207
No Class Precision:	0,966	0,980	0,979
F1 Score:	0,397	0,311	0,313
Accuracy:	0,946	0,854	0,858
Specificity:	0,977	0,864	0,869
Sensitivity:	0,353	0,660	0,646

TableA.4: Only tweet initial scoring random walk iterations with hub weight 0,1 and authority weight 0,1

Experiment:	Iteration=0	Iteration=1	Iteration=2
False Positive:	64	511	403
True Negative:	2786	2339	2447
True Positive:	53	105	95
False Negative:	97	45	55
Yes Class Recall:	0,353	0,700	0,633
No Class Recall:	0,978	0,821	0,859
Yes Class Precision:	0,453	0,170	0,191
No Class Precision:	0,966	0,981	0,978
F1 Score:	0,397	0,274	0,293
Accuracy:	0,946	0,814	0,847
Specificity:	0,977	0,821	0,858
Sensitivity:	0,353	0,700	0,633

TableA.5: Only user initial scoring random walk iterations with hub weight 0,01 and authority weight 0,01 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	53	146	214
True Negative:	2797	2704	2636
True Positive:	53	72	91
False Negative:	97	78	59
Yes Class Recall:	0,353	0,480	0,607
No Class Recall:	0,981	0,949	0,925
Yes Class Precision:	0,500	0,330	0,298
No Class Precision:	0,966	0,972	0,978
F1 Score:	0,414	0,391	0,400
Accuracy:	0,950	0,925	0,909
Specificity:	0,981	0,949	0,925
Sensitivity:	0,353	0,480	0,607

TableA.6: Only user initial scoring random walk iterations with hub weight 0,04 and authority weight 0,04 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	53	345	174
True Negative:	2797	2505	2676
True Positive:	53	105	59
False Negative:	97	45	91
Yes Class Recall:	0,353	0,700	0,393
No Class Recall:	0,981	0,879	0,939
Yes Class Precision:	0,500	0,233	0,253
No Class Precision:	0,966	0,982	0,967
F1 Score:	0,414	0,350	0,308
Accuracy:	0,950	0,870	0,912
Specificity:	0,981	0,879	0,939
Sensitivity:	0,353	0,700	0,393

TableA.7: Only user initial scoring random walk iterations with hub weight 0,07 and authority weight 0,07 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	53	382	363
True Negative:	2797	2468	2487
True Positive:	53	100	100
False Negative:	97	50	50
Yes Class Recall:	0,353	0,667	0,667
No Class Recall:	0,981	0,866	0,873
Yes Class Precision:	0,500	0,207	0,216
No Class Precision:	0,966	0,980	0,980
F1 Score:	0,414	0,316	0,326
Accuracy:	0,950	0,856	0,862
Specificity:	0,981	0,866	0,872
Sensitivity:	0,353	0,667	0,667

TableA.8: Only user initial scoring random walk iterations with hub weight 0,1 and authority weight 0,1 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	53	449	365
True Negative:	2797	2401	2485
True Positive:	53	113	91
False Negative:	97	37	59
Yes Class Recall:	0,353	0,753	0,607
No Class Recall:	0,981	0,842	0,872
Yes Class Precision:	0,500	0,201	0,200
No Class Precision:	0,966	0,985	0,977
F1 Score:	0,414	0,317	0,300
Accuracy:	0,950	0,838	0,859
Specificity:	0,981	0,842	0,872
Sensitivity:	0,353	0,753	0,607

TableA.9: User and tweet hybrid initial scoring random walk iterations with hub weight 0.01 and authority weight 0.01 with user and tweet hybrid initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	53	643	643
True Negative:	2797	2207	2207
True Positive:	53	120	120
False Negative:	97	30	30
Yes Class Recall:	0,353	0,480	0,800
No Class Recall:	0,981	0,949	0,774
Yes Class Precision:	0,500	0,330	0,157
No Class Precision:	0,966	0,972	0,987
F1 Score:	0,414	0,391	0,263
Accuracy:	0,950	0,776	0,776
Specificity:	0,981	0,774	0,774
Sensitivity:	0,353	0,800	0,800

TableA.10: Detailed results of dictionary based experiments

Experiment:	E1	E2	E3	E4	E5	E6	E7
False Positive:	96	132	189	14	166	31	19
True Negative:	2754	2718	2661	2836	2684	2819	2831
True Positive:	62	113	118	54	56	53	51
False Negative:	88	37	32	96	94	97	99
Yes Class Recall:	0,413	0,753	0,787	0,36	0,373	0,353	0,34
No Class Recall:	0,966	0,954	0,934	0,995	0,942	0,989	0,993
Yes Class Precision:	0,392	0,461	0,384	0,794	0,252	0,631	0,729
No Class Precision:	0,969	0,987	0,988	0,967	0,966	0,967	0,966
F1 Score:	0,403	0,571	0,516	0,495	0,301	0,453	0,464
Accuracy:	0,938	0,943	0,926	0,963	0,913	0,957	0,96
Specificity:	0,966	0,953	0,933	0,995	0,942	0,989	0,993
Sensitivity:	0,413	0,753	0,786	0,36	0,373	0,353	0,34

TableA.11: First Dimension Negative Sentiment Based Experiments Results

Experiment:	<= -2	<= -3	<= -4	<= -5
False Positive:	475	166	103	0
True Negative:	2375	2684	2747	2850
True Positive:	72	58	40	7
False Negative:	78	92	110	143
Yes Class Recall:	0,480	0,387	0,267	0,047
No Class Recall:	0,833	0,942	0,964	1,000
Yes Class Precision:	0,132	0,259	0,280	1,000
No Class Precision:	0,968	0,967	0,961	0,952
F1 Score:	0,207	0,310	0,273	0,090
Accuracy:	0,816	0,914	0,929	0,952
Specificity:	0,833	0,942	0,964	1,000
Sensitivity:	0,480	0,390	0,266	0,046

TableA.12: First Dimension Positive Sentiment Based Experiments Results

Experiment:	<= 2	<= 3	<= 4	<= 5
False Positive:	2840	2847	2850	2850
True Negative:	10	3	0	0
True Positive:	148	149	150	150
False Negative:	2	1	0	0
Yes Class Recall:	0,987	0,993	1,000	1,000
No Class Recall:	0,004	0,001	0,000	0,000
Yes Class Precision:	0,050	0,050	0,050	0,050
No Class Precision:	0,833	0,750	0,000	0,000
F1 Score:	0,095	0,095	0,095	0,095
Accuracy:	0,052	0,051	0,050	0,050
Specificity:	0,004	0,001	0,000	0,000
Sensitivity:	0,986	0,993	1,000	1,000

APPENDIX B

EXPERIMENTAL ANALYSIS FOR DIMENSION 2 - SPAM TWEETS

TableB.1: Only tweet initial scoring random walk iterations with hub weight 0,01 and authority weight 0,01

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	139	192	327
True Negative:	2468	2415	2280
True Positive:	224	224	260
False Negative:	169	169	133
Yes Class Recall:	0,569	0,570	0,662
No Class Recall:	0,946	0,926	0,875
Yes Class Precision:	0,615	0,538	0,443
No Class Precision:	0,936	0,935	0,945
F1 Score:	0,591	0,554	0,531
Accuracy:	0,897	0,880	0,847
Specificity:	0,947	0,926	0,875
Sensitivity:	0,570	0,570	0,662

TableB.2: Only tweet initial scoring random walk iterations with hub weight 0,04 and authority weight 0,04

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	139	444	915
True Negative:	2468	2163	1692
True Positive:	224	276	321
False Negative:	169	117	72
Yes Class Recall:	0,569	0,702	0,817
No Class Recall:	0,946	0,830	0,649
Yes Class Precision:	0,615	0,383	0,260
No Class Precision:	0,936	0,949	0,959
F1 Score:	0,591	0,496	0,394
Accuracy:	0,897	0,813	0,671
Specificity:	0,947	0,830	0,649
Sensitivity:	0,570	0,702	0,817

TableB.3: Only tweet initial scoring random walk iterations with hub weight 0,07 and authority weight 0,07

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	139	569	1074
True Negative:	2468	2038	1533
True Positive:	224	292	336
False Negative:	169	101	57
Yes Class Recall:	0,569	0,743	0,855
No Class Recall:	0,946	0,782	0,588
Yes Class Precision:	0,615	0,339	0,238
No Class Precision:	0,936	0,953	0,964
F1 Score:	0,591	0,466	0,373
Accuracy:	0,897	0,777	0,623
Specificity:	0,947	0,782	0,588
Sensitivity:	0,570	0,743	0,855

TableB.4: Only tweet initial scoring random walk iterations with hub weight 0,1 and authority weight 0,1

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	139	662	1135
True Negative:	2468	1945	1472
True Positive:	224	300	324
False Negative:	169	93	69
Yes Class Recall:	0,569	0,763	0,824
No Class Recall:	0,946	0,746	0,565
Yes Class Precision:	0,615	0,312	0,222
No Class Precision:	0,936	0,954	0,955
F1 Score:	0,591	0,443	0,350
Accuracy:	0,897	0,748	0,560
Specificity:	0,947	0,746	0,565
Sensitivity:	0,570	0,763	0,824

TableB.5: Only user initial scoring random walk iterations with hub weight 0,01 and authority weight 0,01 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	1491	2129	2339
True Negative:	1116	478	268
True Positive:	338	370	387
False Negative:	55	23	6
Yes Class Recall:	0,860	0,941	0,985
No Class Recall:	0,428	0,183	0,103
Yes Class Precision:	0,185	0,148	0,142
No Class Precision:	0,953	0,954	0,978
F1 Score:	0,304	0,256	0,248
Accuracy:	0,485	0,283	0,218
Specificity:	0,482	0,183	0,103
Sensitivity:	0,860	0,941	0,985

TableB.6: Only user initial scoring random walk iterations with hub weight 0,04 and authority weight 0,04 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	1491	2447	2500
True Negative:	1116	160	107
True Positive:	338	387	389
False Negative:	55	6	4
Yes Class Recall:	0,860	0,985	0,990
No Class Recall:	0,428	0,061	0,041
Yes Class Precision:	0,185	0,137	0,135
No Class Precision:	0,953	0,964	0,964
F1 Score:	0,304	0,240	0,237
Accuracy:	0,485	0,182	0,165
Specificity:	0,482	0,061	0,041
Sensitivity:	0,860	0,985	0,989

TableB.7: Only user initial scoring random walk iterations with hub weight 0,07 and authority weight 0,07 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	1491	2456	2556
True Negative:	1116	151	51
True Positive:	338	386	390
False Negative:	55	7	3
Yes Class Recall:	0,860	0,982	0,992
No Class Recall:	0,428	0,058	0,020
Yes Class Precision:	0,185	0,136	0,132
No Class Precision:	0,953	0,956	0,944
F1 Score:	0,304	0,239	0,234
Accuracy:	0,485	0,179	0,147
Specificity:	0,482	0,058	0,019
Sensitivity:	0,860	0,982	0,992

TableB.8: Only user initial scoring random walk iterations with hub weight 0,1 and authority weight 0,1 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	1491	2487	2558
True Negative:	1116	120	49
True Positive:	338	392	391
False Negative:	55	1	2
Yes Class Recall:	0,860	0,997	0,995
No Class Recall:	0,428	0,046	0,019
Yes Class Precision:	0,185	0,136	0,133
No Class Precision:	0,953	0,992	0,961
F1 Score:	0,304	0,240	0,234
Accuracy:	0,485	0,170	0,147
Specificity:	0,482	0,046	0,019
Sensitivity:	0,860	0,997	0,995

TableB.9: User and tweet hybrid initial scoring random walk iterations with hub weight 0.01 and authority weight 0.01 with user and tweet hybrid initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	426	616	840
True Negative:	2181	1991	1767
True Positive:	253	294	300
False Negative:	140	99	93
Yes Class Recall:	0,644	0,748	0,763
No Class Recall:	0,837	0,764	0,678
Yes Class Precision:	0,373	0,323	0,263
No Class Precision:	0,940	0,953	0,950
F1 Score:	0,472	0,451	0,391
Accuracy:	0,811	0,762	0,689
Specificity:	0,837	0,764	0,678
Sensitivity:	0,644	0,748	0,763

APPENDIX C

EXPERIMENTAL ANALYSIS FOR DIMENSION 3 - NEWSWORTHINESS

TableC.1: Only tweet initial scoring random walk iterations with hub weight 0,01 and authority weight 0,01

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	351	345	334
True Negative:	581	587	598
True Positive:	1745	1793	1737
False Negative:	323	275	331
Yes Class Recall:	0,843	0,867	0,840
No Class Recall:	0,623	0,630	0,642
Yes Class Precision:	0,832	0,839	0,839
No Class Precision:	0,642	0,681	0,644
F1 Score:	0,837	0,853	0,839
Accuracy:	0,775	0,793	0,778
Specificity:	0,623	0,630	0,641
Sensitivity:	0,844	0,867	0,840

TableC.2: Only tweet initial scoring random walk iterations with hub weight 0,04 and authority weight 0,04

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	351	310	239
True Negative:	581	622	693
True Positive:	1745	1698	1439
False Negative:	323	370	629
Yes Class Recall:	0,843	0,821	0,696
No Class Recall:	0,623	0,667	0,744
Yes Class Precision:	0,832	0,846	0,858
No Class Precision:	0,642	0,627	0,524
F1 Score:	0,837	0,833	0,768
Accuracy:	0,775	0,773	0,710
Specificity:	0,623	0,667	0,743
Sensitivity:	0,844	0,821	0,670

TableC.3: Only tweet initial scoring random walk iterations with hub weight 0,07 and authority weight 0,07

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	351	258	193
True Negative:	581	674	739
True Positive:	1745	1596	1289
False Negative:	323	472	779
Yes Class Recall:	0,843	0,772	0,623
No Class Recall:	0,623	0,723	0,793
Yes Class Precision:	0,832	0,861	0,870
No Class Precision:	0,642	0,588	0,487
F1 Score:	0,837	0,814	0,726
Accuracy:	0,775	0,757	0,676
Specificity:	0,623	0,723	0,792
Sensitivity:	0,844	0,772	0,623

TableC.4: Only tweet initial scoring random walk iterations with hub weight 0,1 and authority weight 0,1

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	351	245	209
True Negative:	581	687	723
True Positive:	1745	1542	1232
False Negative:	323	526	836
Yes Class Recall:	0,843	0,746	0,596
No Class Recall:	0,623	0,737	0,776
Yes Class Precision:	0,832	0,863	0,855
No Class Precision:	0,642	0,566	0,464
F1 Score:	0,837	0,800	0,702
Accuracy:	0,775	0,743	0,652
Specificity:	0,623	0,737	0,776
Sensitivity:	0,844	0,746	0,596

TableC.5: Only user initial scoring random walk iterations with hub weight 0,01 and authority weight 0,01 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	750	841	876
True Negative:	182	91	56
True Positive:	1855	2032	2042
False Negative:	213	36	26
Yes Class Recall:	0,897	0,983	0,987
No Class Recall:	0,195	0,098	0,060
Yes Class Precision:	0,712	0,707	0,700
No Class Precision:	0,461	0,717	0,683
F1 Score:	0,794	0,823	0,819
Accuracy:	0,679	0,708	0,699
Specificity:	0,195	0,098	0,060
Sensitivity:	0,897	0,983	0,987

TableC.6: Only user initial scoring random walk iterations with hub weight 0,04 and authority weight 0,04 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	750	811	886
True Negative:	182	121	46
True Positive:	1855	1944	2010
False Negative:	213	124	58
Yes Class Recall:	0,897	0,940	0,972
No Class Recall:	0,195	0,130	0,049
Yes Class Precision:	0,712	0,706	0,694
No Class Precision:	0,461	0,494	0,442
F1 Score:	0,794	0,806	0,810
Accuracy:	0,679	0,688	0,685
Specificity:	0,195	0,130	0,049
Sensitivity:	0,897	0,940	0,972

TableC.7: Only user initial scoring random walk iterations with hub weight 0,07 and authority weight 0,07 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	750	847	905
True Negative:	182	85	27
True Positive:	1855	1987	2030
False Negative:	213	81	38
Yes Class Recall:	0,897	0,961	0,982
No Class Recall:	0,195	0,091	0,029
Yes Class Precision:	0,712	0,701	0,692
No Class Precision:	0,461	0,512	0,415
F1 Score:	0,794	0,811	0,812
Accuracy:	0,679	0,690	0,686
Specificity:	0,195	0,091	0,029
Sensitivity:	0,897	0,961	0,982

TableC.8: Only user initial scoring random walk iterations with hub weight 0,1 and authority weight 0,1 and only user initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	750	878	887
True Negative:	182	54	45
True Positive:	1855	1960	2021
False Negative:	213	108	47
Yes Class Recall:	0,897	0,948	0,977
No Class Recall:	0,195	0,058	0,048
Yes Class Precision:	0,712	0,691	0,695
No Class Precision:	0,461	0,333	0,489
F1 Score:	0,794	0,799	0,812
Accuracy:	0,679	0,671	0,689
Specificity:	0,195	0,058	0,048
Sensitivity:	0,897	0,948	0,977

TableC.9: User and tweet hybrid initial scoring random walk iterations with hub weight 0.01 and authority weight 0.01 with user and tweet hybrid initial scoring

Experiment:	Iteration=1	Iteration=2	Iteration=3
False Positive:	324	394	641
True Negative:	608	538	291
True Positive:	1758	1811	1985
False Negative:	310	257	83
Yes Class Recall:	0,850	0,876	0,960
No Class Recall:	0,652	0,577	0,312
Yes Class Precision:	0,844	0,821	0,756
No Class Precision:	0,662	0,677	0,778
F1 Score:	0,847	0,848	0,846
Accuracy:	0,789	0,783	0,759
Specificity:	0,652	0,577	0,312
Sensitivity:	0,850	0,876	0,960

APPENDIX D

SUPERVISED LEARNING PHASE RESULTS

TableD.1: First Dimension Supervised Learning Phase Best Results

True Negative:	2771
True Positive:	54
False Positive:	79
False Negative:	96
Yes Class Recall:	0,358
No Class Recall:	0,972
Yes Class Precision:	0,403
No Class Precision:	0,966
F1 Score:	0,380
Accuracy:	0,941
Specificity:	0,972
Sensitivity:	0,360

TableD.2: Second Dimension Supervised Learning Phase Best Results

True Negative:	2468
True Positive:	224
False Positive:	139
False Negative:	169
Yes Class Recall:	0,569
No Class Recall:	0,946
Yes Class Precision:	0,615
No Class Precision:	0,936
F1 Score:	0,591
Accuracy:	0,897
Specificity:	0,946
Sensitivity:	0,569

TableD.3: Third Dimension Supervised Learning Phase Best Results

True Negative:	581
True Positive:	1745
False Positive:	351
False Negative:	323
Yes Class Recall:	0,843
No Class Recall:	0,623
Yes Class Precision:	0,832
No Class Precision:	0,642
F1 Score:	0,837
Accuracy:	0,775
Specificity:	0,623
Sensitivity:	0,844