A Thesis

entitled

Efficient Spam Detection across Online Social Networks

by

Hailu Xu

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the

Master of Science Degree in

Engineering with Concentration on Computer Science

_____

Dr. Weiqing Sun, Committee Chair

_____

Dr. Ahmad Y Javaid, Committee Co-Chair

_____

Dr. Hong Wang, Committee Member

_____

Dr. Amanda Bryant-Friedrich, Dean
College of Graduate Studies

The University of Toledo

August 2016

An Abstract of

Efficient Spam Detection across Online Social Networks

by

Hailu Xu

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the
Master of Engineering Degree in
Computer Science and Engineering

The University of Toledo

August 2016

Online Social Networks (OSNs) have become more and more popular in the whole world recently. People share their personal activities, views, and opinions among different OSNs. Simultaneously, social spam appears more frequently and in various formats throughout popular OSNs. As big data theory receives much more attention, it is expected that OSNs will have more interactions with each other shortly. This would enable a spam link, content or profile attack to easily move from one social network like Twitter to other social networks like Facebook. Therefore, efficient detection of spam has become a significant and popular problem. This paper focuses on spam detection across multiple OSNs by leveraging the knowledge of detecting similar spam within an OSN and using it in different OSNs. We chose Facebook and Twitter for our study targets, considering that they share the most similar features in posts, topics, and user activities, etc. We collected two datasets from them and performed analysis based on our proposed methodology. The results show that detection combined with spam in Facebook show a more than 50% decrease of spam tweets in Twitter, and detection combined with spam of Twitter shows a nearly 71.2% decrease of spam posts in Facebook. This means similar spam of one social network can

significantly facilitate spam detection in other social networks. We proposed a new perspective of spam detection in OSNs.

To my dear parents, thanks for their incomparable love.

To all my friends who give a warm hand to me during this two years' loveable time.

Acknowledgements

I am very grateful to have opportunity studying and learning at the EECS Department, University of Toledo. First, I would like to express my deep and sincere gratitude to my master advisor Dr. Weiqing Sun for his patient guidance, kindness, and strong support during my master program. Also, I want to thanks to my co-advisor, Dr. Ahmad Javaid, who gave me lots of precious suggestions and guidance. I would also like to show my thanks to my committee member Dr. Hong Wang, for his time for my defense and oral speaking.

Second, I want to thanks Dr. Eddie Chou and Dr. Liangbo Hu in Department of Civil Engineering for nearly two years' support. Working in their lab is a fantastic and remarkable experience. Also, I would like to thanks Dr. Kevin Xu, for his valuable suggestions and recommendations. I would like to thanks all the professors I had met at the University of Toledo, thanks for their teaching and guiding.

My sincere thanks then to all my friends that I met with at the University of Toledo, you help me to get an excited short life journey here. Thanks for your help and kindness.

Finally, I would like to thanks my dear family, my mother, and father. Thank you for your supports and understanding no matter under what situations. Without your love and care, I cannot have such fantastic life and experience. Mt best wishes to all of you and thanks you all again.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

API .....Application Programming Interface

DMARC.....Domain-based Message Authentication, Reporting & Conformance
DNS.....Domain Name System

FN .....False Negative
FP .....False Positive
FSD .....Facebook Spam Dataset
FSMD.....Facebook Spam Mixed Dataset

HTML .....HyperText Mark-up Language
HTTP.....Hypertext Transfer Protocol

OSNs.....Online Social Networks

SPADE .....Social-spam Analytics and Detection Framework
SPF.....Sender Policy Framework

TF-IDF .....Term Frequency–Inverse Document Frequency
TN .....True Negative
TP.....True Positive
TSD.....Twitter Spam Dataset
TSMD .....Twitter Spam Mixed Dataset

URL .....Uniform Resource Locator

XML .....Extensible Markup Language

List of Symbols

α .....Ratio of Spam
β .....Ratio of Ham

RP_i.....Revised Probability of *ith* content
P_{1i}.....Probability of *ith* content during the classification of original data sets
P_{2i}.....Probability of *ith* content via the classification combined with outside spam

Chapter 1

Background

1.1 Online Social Networks

1.1.1 Various Social Networks

A social network is a platform for people sharing their activities, interests, background, and real life connections via specific visual computer techniques. A social network consists of a basic background of each user (often a personal profile), his or her social connections, and a variety of additional information such as career and academic backgrounds [1]. Online social networks (OSNs) have become more and more popular in nowadays society, and it would be hard to get rid of them from normal daily life. The very first online social network is an email where people shared and transferred information via different email addresses. Benefiting from the flourish of smartphones, people have multiples choices of various social network applications or Apps. Facebook, Twitter, Snapchat, Tumblr, Instagram, etc. have been a huge part of people's normal life. Figure 1-1 and Figure 1-2 shows the interface of two most popular OSNs, Twitter and Facebook. People share their personal thoughts, activities, arrangements, and information on daily life via different OSNs. At the same time most of the celebrities, athletes, politicians always share their

activities and information via various OSNs. Information in OSNs had been a major role for international organizations and institutions to publish their statements. Also, with the rapid developments of different functioned OSNs, more and more people probably share their activities in the forms of similar posts during different OSNs because of the various scopes of friends and followers. For example, when Justin Bieber plans to publish his new album, he or his company will post the same or similar content on all his social network accounts to notify all followers about the information of this new album. Sometimes, various OSNs design a new function to share their posts on other social networks. People are even able to use a feature on Facebook to automatically publish updates to their Twitter accounts simultaneously [2]. The similar function can also be designed in other social networks, for example, Tumblr users can share the pictures or information to

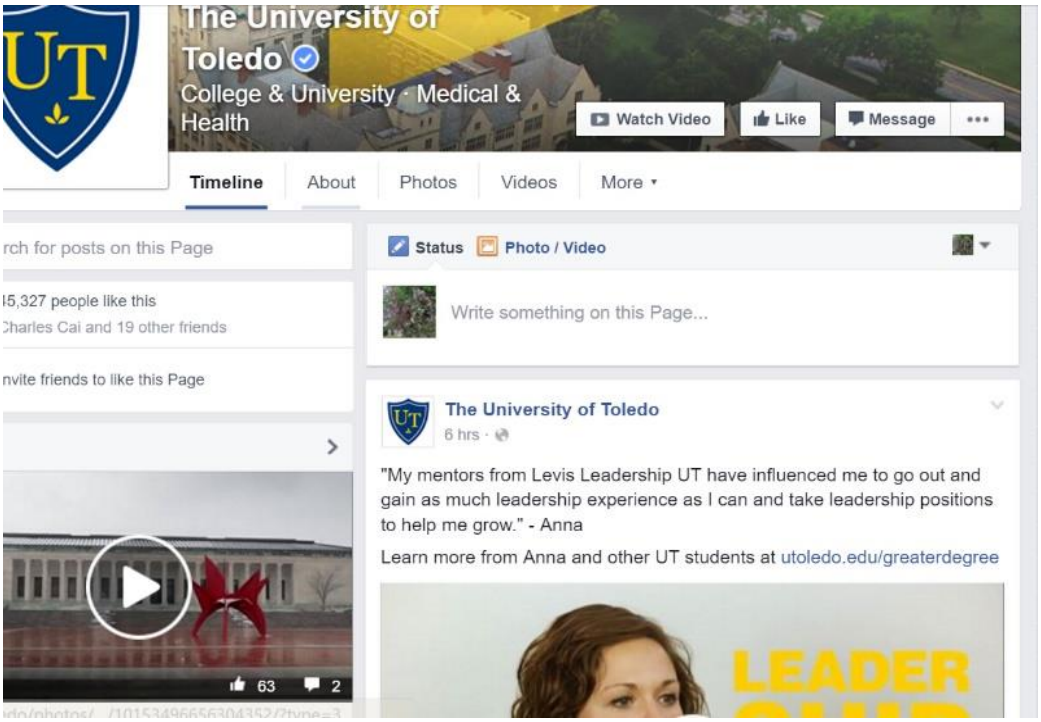

Figure 1-1. Interface of Twitter

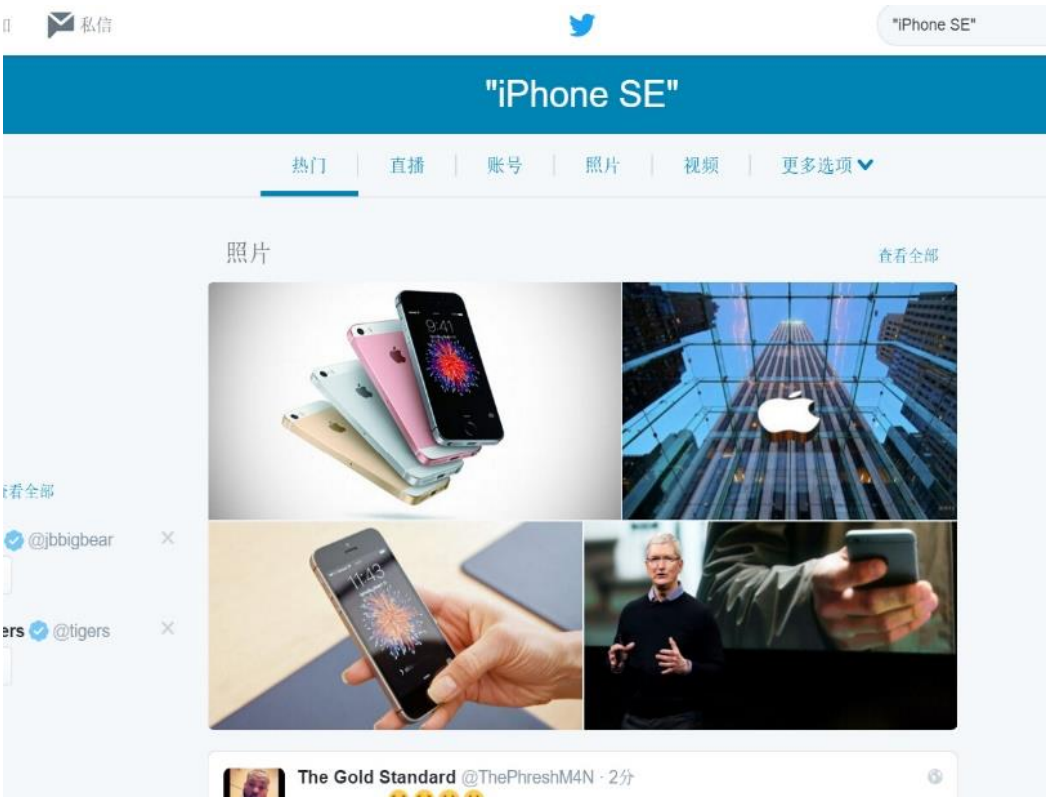

Figure 1-2. Interface of Facebook

Twitter and Facebook accounts. Most of the web pages have various buttons to allow viewers to share the page to various OSNs. All of these make different OSN accounts for one person exhibit high similarities.

1.1.2 Social Networks Security

Designers had created various kinds of policies and technologies to prevent the potential spam activities sincere the birthday of the World Wide Web. As the first global social network, the email had huge amounts of spammers' invasions and attacks. A huge amount of methodologies has been used for email spam detection and social network activities. Unfortunately, high prosperity in OSNs gives rich soils for different kinds of

spams. Spammers who aim to advertise their products or post victim links are more frequently spreading their malicious activities via different OSNs. The information is not private if the information is posted to a social networking site. The more information users show in social networks; the more vulnerable users may become. Even when using high-security settings, the webpages and various activities will leak users' information to spammers. Personal information users share on OSNs could be used to conduct attacks against their associates. The more information shared via social networks, the more likely that someone could impersonate users and his or her friends into sharing personal information, providing access to restricted sites, or downloading malware. Business competitors, predators, and hackers troll social networking sites looking for information or people to target for exploitation [22]. For social network spam, researchers and scientists had developed several popular theories about spam classification and detection. For content spam detection, researchers use TF-IDF to extract word features combined with term frequency [28], retrieve the similarity between different documents. For link classification, people usually choose PageRank to calculate the link relevance by using hyperlinks. Spammers always have their own methods for the link-based ranking or PageRank policy, they use spam links to improve the score of the target page. By TF-IDF, spam words always connect with the contents so that this spam page show more similarity for some queries [28].

Information gleaned from social networking sites may be used to design a specific attack that does not come by way of the social networking site. Reports show that nearly 10% of tweets in Twitter are all spam [3], and Facebook usually blocks 200 million malicious actions every day [4]. In 2008, A market survey showed that at least 83% of the

users of online social networks have received one or more unknown friend request or message. That survey based on the user perception of online social network spam [23]. Even if all companies developed approaches to limit the activities of spammers, spam volume is rapidly growing more than users' actions. Sometimes hackers can modify the code of a social network site, add malicious code into it, always about advertisements and third-party apps ads. On Facebook and Twitter, when log accounts in a work computer, several shortened URLs will point to malicious sites, and that a possible way for personal information leakage. Because it is very easy to retweet a post so that it finally could be seen by hundreds of or thousands of people, Twitter is especially vulnerable to this method [5].

Meanwhile, lacking of social network policy has been a major problem for users. Without authority policy, users cannot have the right ways to protect their interests and privates. Moreover, social network company may have chaos in the protections of users' privates. Sometimes users download more than they needed. In March 2011, Google officially deleted more than 60 applications. Those applications carried illegal or malicious software from Google Android Market [5]. Part of the malicious applications were used to steal the user's private information, then sold to a third, or modify the information or user profile via other devices, or even deleted users' accounts.

Most of the social networks had made several methods prevent different kinds of spam. Users in Facebook and Twitter can click the report spam to notify the employees to delete those posts they think as spam. Moreover, those social networks designed their spam filter system to detect kinds of spam. Though these filters had made a tremendous

improvement in spam detection, we still discover there are huge amounts of spam appear in various OSNs at various time.

1.2 Spam Detection

1.2.1 History of Spam Detection in Email

A large number of spam detection techniques came from the email classification [35]. Those techniques have four parts: individuals' actions, email administrator's automation, email senders' actions and those of researchers and law enforcement officials. For email spam detection, scientists developed several kinds of techniques to prevent spams:

Checking words: spam can be detected via the contents of actual email, either by detecting keywords such as "sexy picture" (content or non-content based).

Lists of sites: From the emails' end user address or consumer ISP, compared with the information in the DNSBLs, which contains the known name of spammers, open relays, and proxy servers, they can be identified as legal or spam.

Several new approaches have been proposed to improve the email system:

Cost-based systems: this solution requires that senders pay some cost to send email, making it prohibitively expensive for spammers who are eager to send large volumes of emails.

DMARC, which stands for "Domain-based Message Authentication, Reporting & Conformance", it standardizes the performance of email authentication using Sender Policy Framework (SPF) and DKIM mechanisms. Channel email is a new proposal which

uses the process of sending an email to restrict anti-spam activities by forcing verification when the first email is received from new contacts.

With the amounts of various social network platforms, spammers have more options and targets to attack. Moreover, that causes the prosperities of spam accounts and malicious posts in OSNs. Lots of scientists and researchers had been focused on this area since a bunch of spam were discovered in OSNs. They also have various kinds of detection methodologies after decades of developments. Many types of research have concentrated on this area to find efficient methods to identify spam and are especially concentrated on the classification of different spam features.

1.2.2 Spam Detections in Current OSNs

X. Hu focused on a content and network information framework for social spammer detection [6]. X. Jin proposed a GAD [7] clustering algorithm to process the challenges about the scalability and real-time detection. B. Markines, C. Cattuto, and F. Menczer focused on six features at the post level, resource level, and user level to specify the spam [8]. H. Gao analyzed spam accounts of social networks to identify the percentage of malicious wall posts, compromised accounts and accounts created for the purpose of spamming [9]. C. Grier, etc. tested the usefulness of URL blacklists to intercept the spreading of Twitter spam via the link feature [11]. M. Bosma proposed a framework combined with user features and spam reports to detect spam [14]. J. Song classified the spams based on the relationships connection features between accounts on Twitter [16]. K. Thomas et al. analyzed different features and behaviors via the largest spam campaigns on Twitter accounts [17]. S. Long designed a new methodology combines word-, topic-, and

user-based features to stem social spam in YouTube [20]. Y. Zhu used a user-activity count matrix to encode the users' social activity in Renren [19]. Basing on spam profile features, K. Lee proposed a honeypot-based approach for spam detection in MySpace and Twitter [18]. K. Thomas etc. found that it can spare 70% of victims by preventing the spread of compromise in 24 hours [10]. J. Caverlee, L. Liu, and S. Webb proposed a reputation-based trust aggregation framework to test spam in MySpace [15].

However, prior research mainly concentrated on spam detection in one particular social network like Twitter or Facebook, and they paid less attention a popular phenomenon, that is, the more and more similarities between various social networks. Several researchers have focused in this area for a period. De Wang, D. Irani, and C. Pu designed a framework called "SPADE" to deal with spam in different social networks and webs via one framework [12]. It can specify various types of spam-like links or contents in various OSNs via particular models.

As activities between different OSNs establish more connections, OSNs will develop more interactions with each other. Therefore, when one spam link, content or spammer attacks one social network, it is possible to appear on other social networks with similar actions. Therefore, if different social networks' spam detection models have the ability to communicate with each other, it will greatly decrease the spam actions. Our research mainly focuses on combining spam in one social network to reveal and intercept new similar spam that may appear on other social networks. We analyzed behaviors and features of spam in various OSNs, and then use the similar features to facilitate the spam detection in other OSNs.

Chapter 2

Methodology

2.1 Data Sets

2.1.1 Social Networks APIs

An application programming interface (API) is a set of procedures, protocols, and tools; it is used for construct various applications and software. Social network platforms offer APIs to users to develop various new web applications. That will benefit its programming structure for outside groups to utilize and create new features to their websites [21]. An API usually consist of an operating system, a web-based system, or a database tool, and always based on a specific programming language. It is useful for developing applications for the different system. APIs can work as the GUI components, or to access computer hardware or database like the hard disk driver. Through various APIs, third parties and researchers have access to the instant data, user activities, celebrities' actions and the most popular topics in the world. In this section, we will introduce the background information about Facebook API and Twitter API, and the datasets collected during the research and then classify research goal before we analyze the datasets.

2.1.1.1 Twitter API

Twitter is one of the most popular OSN in the world; Twitter users can use *tweets*, *hashtags* or *mentions* for sharing information and activities with their followers. The *tweets* allow users to share a link or update with words up to maximum 140 characters; by using *@mentions* users can directly address anyone they want; *hashtags* allow users to update following with several keywords or group activities, and the post will begin with a "#" character. A Twitter user can follow another user by clicking *follow* button. By following the people they interested, he or she will receive their interested people's tweets on their page. The user who had been followed, if she or he wants, can follow back by the same functions. When a user likes someone's tweet, he can decide to retweet it or not. As a result, all her followers can see that message. In default setting, users' profiles are all public inTwitter, but a user could protect his or her profile by security settings if they desire. With that, if anyone wants to follow that user, he must need that user's permission.

Twitter API [24] allow certificated users to search information via different ways: There are four main "objects" that users could use from the API: *Tweets*, *Users*, *Entities* (also *Entities in Objects*), and *Places*. Meanwhile, the API should also include with *Twitter IDs* or *Place Attributes*. Most of the people use *oAuth* to get the access to Twitter API. The request of a person's signature is determined by the identity of their application, in addition to the identity of the user's access to the identity that is granted by the end user, and the access token of the user's access token is represented by the interface. By using the *keyword*, API can reach information around this keyword via the whole world tweets; by

locations users can search other users' posts and information, mainly focusing on one city or place; by *following* users can search all tweets, retweets, and replies which are about one user, etc.

Twitter API is not free for users to do whatever they want to. Twitter sets a certain limit in its API to usefulness prevent the damage to the bandwidth from the killer or spammers.

For users, it only allows a maximum of 180 requests each 15 minutes. While this restriction only applies to getting (instead of POST) request, but former experience has shown that it is an excellent rule. If users exceed this limit, the document produced by the REST call will tell users about this. So, no matter what reasons, when users call the Twitter REST API over 180 queries per fifteen minutes, Twitter will response whitelisting to them.

Another limitation is the case that regardless of how many page count or parameters, it only returns up to 3200 states. Also, Twitter only requested but not mandatorily own other restrictions. For example, Twitter is recommended to use page attributes, deprecated count property. As another example, it recommended that the results should be saved to a local cache not to repeat the request with the same state.

In general, there are two different forms of the HTTP request, POST, and GET. These two forms also invoke Twitter API.

Simply put, it forwards the POST and GET requests from clients to the original API address, and will return the HTTP header and the contents back to the client, that fulfills all the features of the original Twitter API. For the client, in addition to providing an option of alternative configuration API address, they do not need to change any of the code. For the following scenarios, usually the most commonly method researchers use to see some

information is to access twitter.com directly to see the Friends list, in fact, it calls the GET request.

2.1.1.2 Facebook API

Facebook is the most popular OSN in the whole world; it has more than one billion active users in the world and billions of posts daily compose the largest online society in the virtual world. The most popular and successful feature of Facebook is its platforms including *wall posts*, *fan pages,* and *tags*. These allow normal users to interact with their favorite celebrities and friends by sharing information and activities. They also can use *tags* to address their friends and events during their post (similar to the *@mention* in Twitter). Usually, user profiles are private and others cannot access or view their profiles if they do not make connections with each other. When user A wants to add user B's as friend, first, the webpage will send user B an asking request, if B knows about use A, then this connection can be built. After B accept the request, user A will be in the friend list of user B. However, the user's friends list in Facebook is different from their actual friends' relationships in the real world. Most of situations, Facebook users finally will accept the friendship request from people they don't know, but in a real world, friendship would require more time and scrutiny [26].

The most commonly used API for the research in Facebook's access is graphics API [25], which is a 'social graph' concept named - composition information on Facebook: node is basically stuff, like a photo, a user, a page, a comment; edge - such as a photo or a comment that something between them; field information about those, such as a user's

location, a person's birthday, or the page’s name. The graphics API is based on HTTP, so it has an HTTP Library of any language, such as cURL, urllib. Researchers can also use graphics API directly in the browser. Each node has a unique ID that is utilized to access it through a graph API. Graphical API can access to multiple versions at any one time. Each version has some core areas and edge operations. Facebook provides these core API, which can be modified in the version from the release of at least two years.

Facebook API allow authenticated users access to a variety of accounts in the network, but it has several limitations: people can access the public accounts and the open group without external authority, but the collection of personal account information. They cannot have any private information account permissions. API Facebook allows users to use feed, position and other parameters to access the data. The core of Facebook Platform API is the Graph API, which lets users read and write data from Facebook. Facebook also has the Old Rest API. The new Graph API change the paradigm oriented approach from the way of reading and writing data from Facebook to a new way, that is, using the object (for example, user profiles, friends, posts, photos, and so on) and relationships or connections between each other. This approach simplifies the Facebook API, making it more consistent when handling objects. Note that although the Graph API is the preferred and most popular Facebook API nowadays, the Old REST API is still active and supported. Graph and the REST API are both suitable for mobile applications (including native and mobile web applications), which through the use of web content with mobile WebViews in a native application.

Graph API object is assigned a unique ID, it is easier to use a URL to access it, this URL can be further defined to address a particular object or connection. For example, a page with the following connections: feed or wall, photos, notes, posts, members, and so on.

By using Graph API, users can retrieve objects, delete or post objects. Users can search, update objects, filter search results, and even automatically match or discover the connection and relationship between particular objects.

In default, the application owns access to users' public data. To access private data, the application must first request the user's permissions (called extended permission). Facebook defines some rights; more information can be found on the official Extended Permissions page on Facebook.

Most of the current social, microblogging sites use OAuth authentication standard; Facebook has no exception. For user login, it needs to enter account password on the Facebook web page, after the login is successful then it will be redirected to the page to get the Token, users need to use Token to visit API validity. That application only needs to get the user's Token; the whole process will not the reach user's account and password. Even if users get the account number is useless, API only recognizes the Token. Compared with HTTP Basic Authentication (HTTP Header increase over Bae64 account and password), the entire practice is more complicated but safer. Basic Authentication era developer can be arbitrary, do a variety of applications. The most troublesome after OAuth is to get verification, especially for desktop or mobile applications, they typically embed a browser control in a form to complete it.

2.1.1.3 Theory of Data Sets

A set of data (or dataset) is a collection of data. The most commonly used data sets correspond to a single database table, or a single statistical data matrix, where each column in the table on behalf of the contents of a particular variable, each row corresponds to a given member of the data in the problem set. The data for each variable, such as height and weight of the objects set list of values, each member of the data set. Each value is referred to a reference. The dataset may include one or more components of the data corresponding to the number of rows.

A collection of tables is closely related to the term data refers to data sets can also be used more loosely, especially for a particular experiment or event. The instrument by space agencies and space probe experiment collected data sets are examples of this type. In the simplest case, it only has one variable, then it consists of the value of a column composed, often described as a list. Usual order is not important; then this set value may be considered multiple sets, instead of an (in order) list.

The value may be a number, such as real integer numbers, for example on behalf of a person's height in centimeters, but may also be symbolic data (i.e., not including digital), for instance, on behalf of a person's race. More generally, the value can be of any type is described as a certain degree of measurement. For each variable, usually, all values are similar. However, it may be "missing value", which should be noted in some way.

Data sets can be divided into a typed dataset and untyped datasets.

Typed dataset: Typed dataset is first got from the Data Set class. It then uses the XML schema file (.xsd file) information to create a new class. Schema information (tables,

columns, etc.) generates and compiles the data established for this new class. It can be directly referenced tables and columns by name in the VS.NET can IntelliSense type elements.

Non-typed dataset: This dataset has no related architecture of the building. Moreover, same as a untyped dataset, typed datasets also contain tables, columns, etc., but they are only sets of disclosure. They need to use Tables reference column.

2.1.2 Data Collection

Suitable labeled datasets are crucial for the whole classification process. As our designed research target, we aim to use different datasets which are gathered from different OSNs but share some similarities in contents or topics. We look forward to analyzing the behaviors and activities of similar groups in different OSNs where they share the most overlapped members. That means that we need to get the datasets from different OSNs but should have similar keywords or activities. At the same time all data should appear in OSNs near the same period, considering that if data is collected from a different period, the original classification frameworks of OSNs themselves will delete most of the spam. So all these conditions make harder to find fitted datasets in other prior research works. We notice some former researchers introduced or published several labeled datasets from Twitter, Myspace, and other OSNs, but unfortunately they were collected in different times and separate topics. Also, most of them are in different formations, so they are all not suitable for our research targets. As most of the spammers' purpose is advertising products and

fishing users to click the malicious links. So this research's data were collected data by ourselves.

The different policies of APIs in different social networks make it difficult to get most appropriate data from various OSNs. For example, Twitter API allows users to search by keywords, location, and so on, but when searching the data, we always get the data from all over the world. It is hard to select only one special group or account in the limited spatial area. Also, the APIs of Facebook only allow users to search and collect data from public accounts, public groups, and private accounts which they allow others access to. As formerly mentioned, if two groups in different OSNs share most of the same members, there must be more similar posts or activities in these groups, and also more similar spam appear. So different functions of APIs limit several aspects of datasets this research's prior designed. We finally decide to collect data via same keyword or topic to make this data as related as possible. Also, we consider more spam appear in OSNs connected with celebrities and popular activities. We collected one dataset via API of Twitter in a keyword ("Taylor Swift") from June 2015 to August 2015, and gathered data of one open group ("World of Taylor Swift") through API of Facebook. Then we labeled these data and normalized them into two datasets.

2.1.2.1 Twitter Data Collection

Twitter API is a part of Twitter REST API. Researchers should use a Streaming API if they want to match completeness. Here are the details about build query and search related data [24]:

First, users need to run the search on twitter.com.search, then check and copy the URL which is loaded.

Then use “search/tweets.json” to replace “/search”. Then it will get: a new URL which end with “q=%40twitterapi”

Final. Execute that new URL link.

There are also several parameters for the GET search in Twitter API, like Result Type, Geolocalization, Language, and Iterating in a result set, those different parameters decide the data users want to search. Remind that Twitter API has its limits [24], access tokens only allow 180 requests/queries per 15 minutes for users.

2.1.2.2 Facebook Data Collection

The new updates on the Facebook search can help users to find friends in a specific city, a particular topic, restaurants, and a lot more photos. In fact, some users need to search and map the options for searching public posts.

Here are the steps for Facebook API search [25]:

The first step: Access Graphics Explorer API page. To perform a search for public duties, it will use the graphical API resource manager tool provided by the developer section on the Facebook. Users can find it on the tool page.

Second step: set the search query. By default, "GET" fields will show "me? Field = ID, the name is ", then it needs to change it to" search? q = "to perform basic search.

The third step: change the version number. Users need to change the "API version" of "V2.0" to the old "V1.0" search function to get the job. If users do not change it, it will get an error.

The fourth step: Get an Access Token. To use the Graph API Explorer, users need an access token. So users need to click on the "Get Access Token" to get one. After the login page appears, then select "Okay" to give the Graph API Explorer app access to public profile (name, profile picture, age range, gender, language, country, etc.) and friends list (which anyone can access using this friend list hack). Users will see a very large alphanumeric code in the "Access Token" field. Via these large access token field, this is what allows users to do the search that the research is about to perform. Figure 2-(a) and (b) show more details about getting the access.

Figure 2-1. API of Facebook

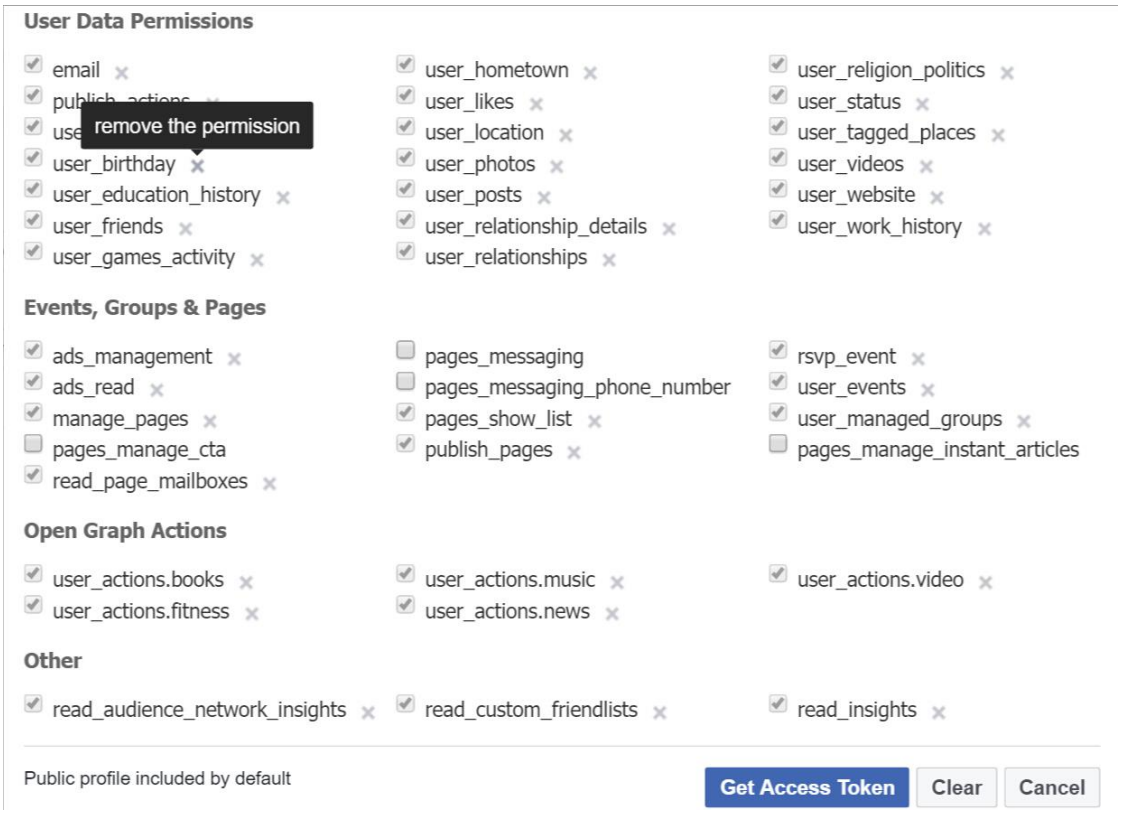

Figure 2-2. API access permission of Facebook

2.1.3 Analysis of Data Sets

We labeled the data and specified them into spam and ham based on the pointed URL links, actual contents in the tweets or posts, and the official identifications to the accounts. Specifically, when one tweet or post contains the URL link which points to fishing, unrelated content advertising, or porn web page, we defined it as spam. Also, it is spam when the official network shows this post or account as an illegal activity or account. All data were saved in CSV file and consisted of contents and categories (spam or ham). Here are more details about these two datasets:

Twitter Spam Dataset (**TSD**): We collected dataset whose keyword was set as “Taylor Swift” in Twitter from June 2015 to August 2015. After our labeling and normalizing, we got this dataset that consists of 1937 spam tweets and 10942 ham tweets.

Facebook Spam Dataset (**FSD**): We collected data from the open public group on Facebook, which was named as “World of Taylor Swift” from July 2015 to August 2015. We labeled and normalized them into one dataset that contains 1338 spam posts and 9285 ham posts.

We analyzed these two datasets via various perspectives. For TSD, in 1937 spam tweets, there are 1473 (75.6%) spam tweets contained in the URL links, 464 (24.4%) spam tweets only contained in words. For the 10942 ham tweets, there are 6877 (62.9%) tweets which consist of URL links and words, and 4065 (37.1%) consist of only words. For the spam posts of FSD, 340 (32.8%) spam posts consist of URL links and words, 998 (67.2%) of spam posts consist of words. For ham posts, 8514 (95.1%) consist of URL links and 771 (4.9%) only consist of words.

Figure 2-3 and 2-4 show the original date format in the CSV file of TSD and FSD. Table 2.1 and 2.2 show the top 20 word features in the spam of Twitter and Facebook. After omitting all nonsense and useless nouns, verbs, and pronouns, they share some similarities in these words that interacted with products advertising, photo, and videos, etc. Table 2.3 shows the most frequent tweets and posts in the spam of Twitter and Facebook. After the analysis, we discovered that most spam tweets and posts consisted of several words and links that point to fishing pages or products advertisements, or they consisted of several useless or nonsense words. These show that spam posts and tweets are always in similar formats.

```
tweets,class
'HaHaHaHaHaaHaHaHaHaHah T-RASH.',Spam
'RT @gima2327: Watch Taylor Swift and #Sam #Hunt http://t.co/UGR8kz1gUJ http://t.co/goGZ8o063i',Spam
'LMFAAAAAAAAAAAA000000000000 https://t.co/S5txRK8dXd',Spam
'QUE https://t.co/eekWTZGTu3',Spam
'I am so done ?????? https://t.co/SafvEoPx8w',Spam
'#NOWPLAYING Sam Smith-Like I Can. Text iPad to 313131 when you hear us play #TaylorSwift #BadBlood & possibly win an i
'RT @Polaroid: Omfg Taylor Swift is such a bitch.',Spam
'look here by using these pictures youre proving @nick https://t.co/1r8xrfZBOe',Spam
'They left . all alone the big trampolini ... \':',Spam
'SAYS COLOR ME.',Spam
'RT @lunarnomad: Taylor Swift is a feminist yet cant grasp how race & gender intersect all while profiting from black c
'Photographers and her fans make nice over photo contract revisions - http://t.co/dr29vAegAA',Spam
'@molliewayout she is so sexy https://t.co/LhOtolNB0q',Spam
'RT @LYEFRDS: literally the fucking captain of white feminism',Spam
'RT @WORLDSTARC0MEDY: the fact that shes going that Hard to taylor swift just amazing?? https://t.co/VE3Ayuz3IQ',Spam
'Taylor Swift 1989 Tour VIP Memorabilia Tickets- Full read by eBay http://t.co/oHSladKy36 http://t.co/7PqlvTTkMt',Spam
'RT @01362: Watch Taylor Swift and Sam Hunts Steamy Take Your Time Duet | Sam Hunt http://t.co/r9Hkk2e5qw',Spam
'RT @LYSHREGUI: literally the fucking captain of white feminism',Spam
'Dressing-Like-Cher-From-Clueless http://t.co/woslo00uQ5',Spam
'A little louder for the people in the back. https://t.co/XKRoRwnXdo',Spam
'#NOWPLAYING Tori Kelly-Should??ve Been Us. Text iPad to 313131 when you hear us play #TaylorSwift #BadBlood & possibly
'sipping this tea ?? ? https://t.co/bt3EEIWE85',Spam
```

Figure 2-3. Data format in Twitter Spam Dataset

```
content,class
'PrisLovatoJonas You Shouldd(:',Spam
'". . Hi im back :-) :-)and my free Data on my cellphone our  already back now <3#socanwechat?? "',Spam
'Ada Eralda posts porn. Pls. Block her admins. anyone? #swiftyLove ',Spam
'"Bakit marami nagsasabiNa mukha daw bakla c tay !)..#PakiExplain..#grrrrrrrrrrrrrr "',Spam
'"Birthday Game #2Comment your Birthday and Find your Twin "',Spam
'"Fansign ALERT!20 Names I don\'t accept ""ME""www.facebook.com/fansigning "',Spam
'"Hey everyonePlss follow me on instagram',Spam
'"http://bit.ly/1K9NiHPClick Hear Photo and Lock Fail pics "',Spam
'"Hy ties! Follow my akun on instagram : natasyaherliza_ follow ties? After you follow aim follback "',Spam
'"Must watch Boy with a female voicehttps://www.facebook.com/video.php?v=837642082986451 "',Spam
'People PLEASE jjoin my new group RT @ https://www.facebook.com/groups/470794936435559/470795746435478/?comm
'"TrickAddicts. ComFree CallingFree RechargeFree InternetFree Facebook TricksFree WhatsApp TricksPaytm tricl
'http://virallisty.rocks/sgcoo-amy-actually-during-was-john-jokes-her-cena-sex-their-inside-schumer-that-sce
'"https://www.facebook.com/photo.php?fbid=129490577404556&set=gm.10153646896364766&type=3  @Current News- "'
'  girl cute ....dance sex......verry nice',Spam
'  I thought this group is Only for swifties and to post only about Queen swift... what the hell are u talki
'"https://www.facebook.com/photo.php?fbid=491433594364418&set=gm.10153548645709766&type=3 Photos from Ash Is
'http://freewatchingcutegirldancinginroom.blogspot.com/2015/07/super-model-sleepy-style-in-publicit-is.html
'  OMG she so sexy :3 :3',Spam
'http://bit.ly/1CTnX2N Guys Love See Through Sexy Yoga Pants Guys Love See Through Sexy Yoga Pants',Spam
'http://bit.ly/1KWXy6m A Husband Divorced His Wife After Looking Closer At This Photo He Took Of Her A Husba
'"https://www.facebook.com/photo.php?fbid=883520548387928&set=gm.10153328559119766&type=3  Hi taylor swift,
```

Figure 2-4. Data format in Facebook Spam Dataset

Table 2.1. Top 20 word features in the spam of Twitter

Top 20 Word Features of Spam Tweets in Twitter			
Freq.	***Word***	***Freq.***	***Word***
1568	https	117	Now playing
411	hard	117	Text
394	fact	116	Fuck
392	going	81	WorldOfDancing
237	wow	77	Watch
234	iPad	73	Hunt
142	amazing	62	Picture
123	win	45	Photo
120	sam	42	eBay
117	hear	42	Tour

Table 2.2. Top 20 word features in the spam of Facebook

Top 20 Word Features of Spam Posts in Facebook			
Freq.	***Word***	***Freq.***	***Word***
374	http(s)	24	fuck
250	money	24	Internet
80	sex	24	marketing
72	mobi	24	fast
50	sexy	22	photo
40	Justin bieber	20	cutshare
36	online	20	video
30	business	20	pregnant
30	new	18	blog
30	free	18	model

Table 2.3. Top 5 spam tweets or posts on Twitter and Facebook

Top 5 spam tweets on Twitter
@WorldOfDancing: the fact that shes going that Hard to taylor swift just wow https://t.co/8dmsaoYOPG
@gima2327: Watch Taylor Swift and #Sam #Hunt http://t.co/UGR8kz1gUJ http://t.co/qoGZ8o063i
@Ratchet: Unseen photo of Ed Sheeran and Taylor Swift http://t.co/LKT9CHaabe
@LYEFRDS: literally the fucking captain of white feminism
@Nasarfa: HaHaHaHaKiakiakiaa https://t.co/mfe9pfA9Mb
Top 5 spam posts in Facebook
"Breaking news: Taylor swift f**k with me ....and now doctor say she is pregnant !!!!"
"Private video of Selena Gomez is revealed on facebook woah woa http://www.hotnews.sexyi.am/news/"
"Hi, Visit (y) www. LoveLiker .com (yThis Site Gives 100+ Likes (y50+ Comments On Post <3 and 50+ Followers: So Visit Fast: I am promoting it. :D"
http://freewatchingcutegirldancinginroom.blogspot.com/2015/07/super-model-sleepy-style-in-publicit-is.html Cute Girl Dancing: Super Model sleepy style in public. It is for advertising new car model. It is so interesting. It is not porn or sexual Video
http://myurl.cz/7zux3 HOT HOT we talk about sex :v :v

2.2 Detection Framework

2.2.1 Detection Strategy

Here first to introduce the basic terms about spam classification approach [28]:

Classification: Classification means researchers first to build a model for a group of classes or concepts, then use the model to predict class labels for test data. For example, to classify whether an email is email spam, web page is web spam.

Prediction: Prediction focuses on the continuous-valued functions of models researchers created. For example, scientists use the prepared models to forecast the economic growth in the next year.

Classification and prediction are a two-step process, which means, model construction and model applications. Model construction means scientists first need to introduce a set of predefined classes, which called training dataset. Training dataset consists of tuples for building a model, and each tuple or sample belongs to a predefined class. At the same time, researchers need to make the classification rules, classification models, decision trees, decision rules, or math formulae, etc. Model application means to classify those unseen objects: researchers need to use an independent test data set to estimate the accuracy of the model, then use the model to classify unknown class labels. The training dataset needs to use some features to make a further application. As former researchers' experience, most of the features are web page top domains, languages, some words (body and title), average word length, anchor words, visibility of content, repeating keywords, the most common keywords, n-gram likelihood and so on.

We suppose to explore the influence of spam in one OSN to another; we do not aim to show how great performance of detection only around one dataset. So we chose 10% original data to do the training work so that it can maintain the maximum independence and testability of posts in one social network, at the same time, it is more intuitional and beneficial to show the influence of spam related with same topics in other social network to the spam detection in that social network. The strategy of process are as follows:

A. We first split TSD and FSD separately into training and test datasets, use training datasets to train the various classifiers, and then use classifiers to check the test datasets. We then get the original classified results of Twitter and Facebook Spam Dataset.

B. To show the influence of Facebook spam posts in Twitter spam tweets classification, this research combines spam of Facebook into a Twitter training set; we then use the newly trained classifiers to test the remaining dataset.

C. After step 2, we then do the same procedure in the Facebook training dataset, and then apply the new training process to verify the original test dataset.

D. Finally, we combine the results of classifications on the above two social networks.

We combined *FilteredClassifier* to train and test with various classify algorithms and used *StringToWordVector* to process natural language in Weka [13]. We also use precision, F1-Measure as criteria to evaluate the classification performance. The relations of the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) are shown in Table 2.4. True positive: Facebook users correctly identified as Facebook users; False positive: Twitter users incorrectly identified as Facebook users; True negative: Twitter

users correctly identified as Twitter users; False negative: Facebook users incorrectly identified as Twitter users [27]. In general, Positive means identified, and negative means rejected. Therefore: True positive means correctly identified; False positive means incorrectly identified; True negative means correctly rejected; False negative means incorrectly rejected.

In information retrieval, precision means the positive predictive value, and recall means sensitivity. For the performance of positive class, the F1-Measure can be a useful measure. The F1-Measure is also a balance measure of precision and recall. Here are the definitions of Precision of spam, F1-Measure, Recall of spam and accuracy based on above terms:

Table 2.4. Relationships of TP, TN, FP and FN

Actual label	**Predicted label**	
	Spam	Ham
Spam	True Positive	False Negative
Ham	False Positive	True Negative

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} - \text{Measure} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

At here, I will briefly to introduce basic theories about our classification algorithms:

Naïve Bayes [29]: Naive Bayes is a simple technique for classifiers: drawn the class labels from some finite data, then use the class labels to notify the problem instances, represent those as feature values. Naïve Bayes classifiers assume that a special feature is isolated of the value of any other features. For example, fruit may be considered to be an orange if it is yellow, round, and about 8 cm in diameter. Each of those features will independently contribute to the probability that it is an orange via naïve Bayes classifier, no matter about any relations between the diameter features, color features, or roundness features.

J-48 [30]: Also named as C4.5, which is an algorithm used to generate a decision tree. This algorithm was developed by Ross Quinlan. By using the information entropy features, after training work via the train data set, C4.5 creates decision trees. Classified samples compose as a training dataset. Each sample is a p-dimensional vector. This algorithm has several basic theories: All the samples in the list belong to the same class, when choosing that class, it simply only creates a leaf node. In addition, no information gain will be provided by features.

Random Tree [31]: The random tree is created by a stochastic process, and also it belongs to tree theory. Random trees have various types, such as uniform spanning tree, Rapidly-exploring tree, Random minimal spanning tree, Brownian tree, Random forest, Random binary tree, Random recursive tree, randomized binary search tree, Branching process[31].

Random Forest [32]: Random forests are a concept of the general technology of random decision forests. Random forest is a learning algorithm for classification, regression and other processes. It will create a number of decision trees in the training time and output classes, those classes are the model of the class (classification) or average prediction (regression) of individual trees. This method combines the concept of Braman's "baging" and the random features selection. Ho, Amit, and German separately presented the theories of bagging and random selection, they also established a set of control variance decision tree.

Bayes Net [33]: also known as the reliability of the network, is an extension of the Bayes method. Bayes Net is a currently uncertain knowledge expression and reasoning in the field of one of the most effective theoretical model. Since 1988 published by the Pearl, it has become the focus of research in recent years. A Bayesian network is a directed acyclic graph, by representing variable nodes and connecting these nodes to the form of an edge. Nodes represent random variables; nodes have to edge represents the inter-node relationships with each other (the child nodes the directedness by the parent node), with conditional probability expression intensity, no parent node with prior probability to convey the information. Node variables can be abstracted from any problem, such as test values, observations, comments, etc.. Bayes Net is also applicable to the expression and analysis of uncertainty and probability of the event. Also, it applies to a conditional reliance on a variety of control factors of decision-making, imprecise or uncertain knowledge or information.

Logistic [34]: also known as regression analysis. It is mainly used in the epidemiology, and more commonly used in the case is to explore the risk factors of a disease, according to the risk factors, which can be used to predict the probability of occurrence of a disease, and so on. It is also used for prediction. If the logistic regression model has been established, it can be based on the model, predict the case of different independent variables, the probability of occurrence of a disease or some situation. Also, it can be used for discrimination, actually was somewhat similar to forecast. When based on the logistic model, it can be utilized to judge someone belonging to a certain disease or belonging to a probability of, also is to take a look at this person has the possibility is a disease.

2.2.2 Contributions

Our research's major contributions are as follows:

1. We propose a new perspective of the spam detection in online social networks. Traditional detection methods are focused on only one social network. However, our work concentrates on spam similarities in different OSNs to analyze and detect such activities.
2. We collected two datasets from Twitter and Facebook through their APIs, each of them contains spam and non-spam contents.

Chapter 3

Classification Results

3.1 Twitter Original Classifications

All classifications are done in Weka [13], which has been one of the standard tools in data mining and machine learning. It was designed by computer professors and scientists at the University of Waikato. It is a free software for research in machine learning and data mining. Weka is a workbench that contains a number of visualization tools and classifiers for data analysis, data classification, and prediction. Also, the graphical user interfaces help users to easily access these functions. Figure 3-1 shows the interface of Weka. Weka is a large group of machine learning algorithms and at the most time used for data mining tasks [13]. Users can directly to use the algorithms to a dataset, or those classifiers can be called from Java or another programming language. Weka can be used for data pre-processing, data classification, model constructions, regression, data analysis, clustering, model analysis, decision rules, and visualization [13]. It contains various classifications and clustering algorithms like Naïve Bayes, J-48, Random Tree, Random Forest, etc.

We use the training datasets to train classification algorithms and then use the algorithms to test the remaining test datasets. We use several different cluster algorithms to detect the spams and calculate the accuracies; we also use one of the most efficient classifiers to classify the test dataset in various percentage. The results of TSD are given in Table 4. Figure 3-2 shows the interface of Weka after the classification. Figure 3-3 shows the details of J48 tree construction when use J48 as a classifier to do the classification. Figure 3-4 shows the visible margin curve when use J48 do the classification in the Twitter data set. Figure 3-5 shows the classification errors after J48 classifier for the Twitter data set.

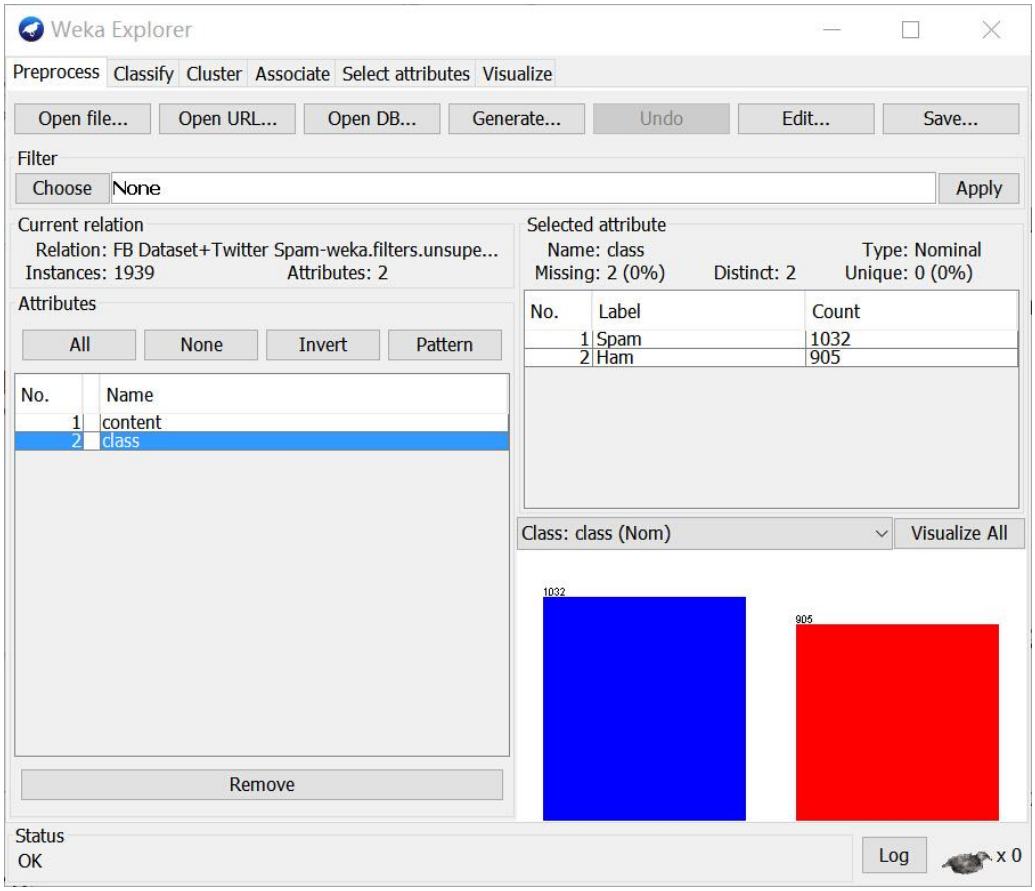

Figure 3-1. Interface of Weka tool

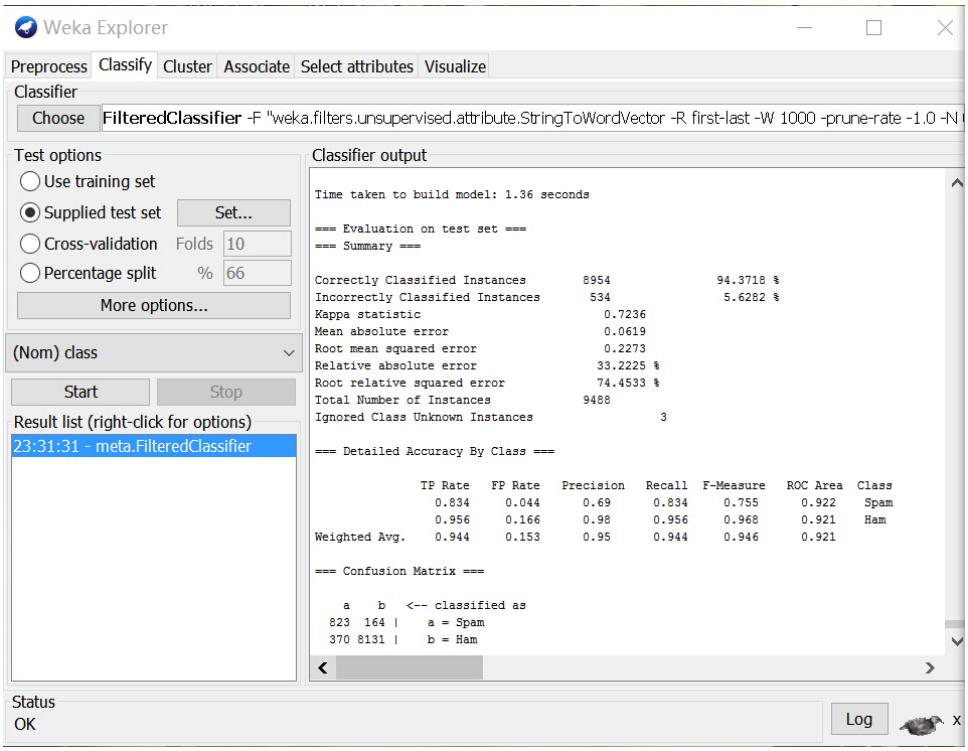

Figure 3-2. Interface of Weka after classification

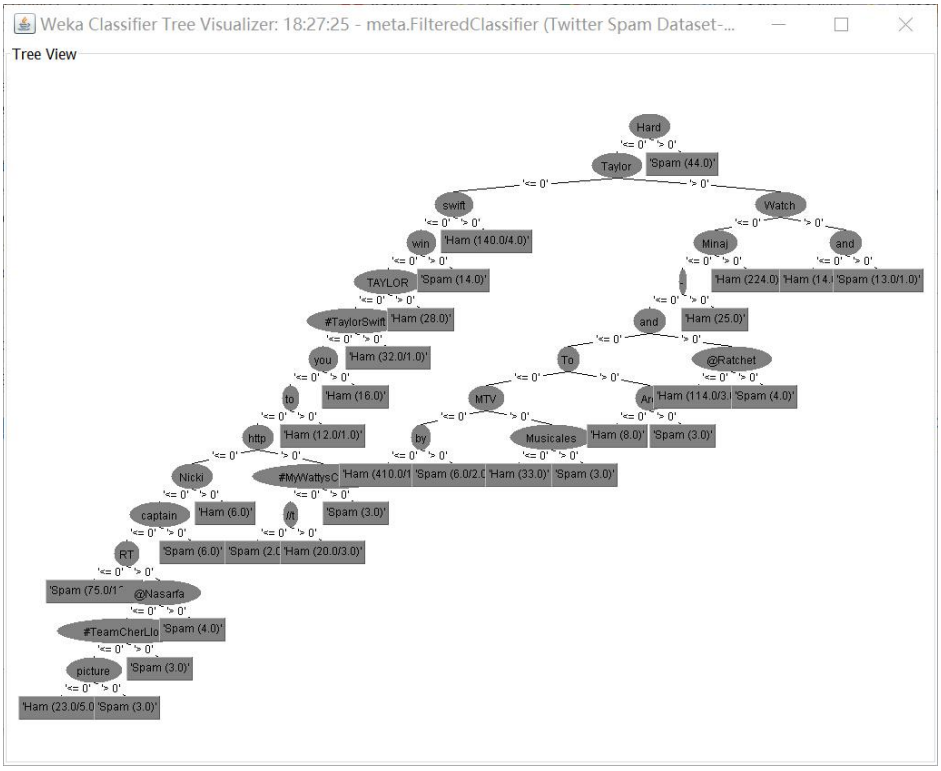

Figure 3-3. Details of J48 Tree via Weka classification

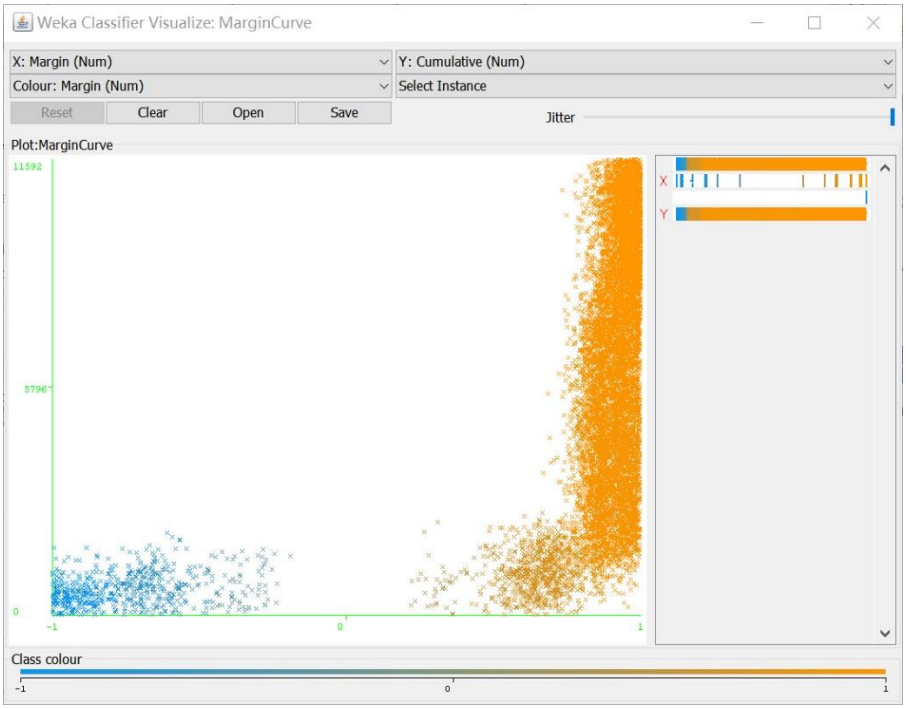

Figure 3-4. Margin Curve of J48 Tree via Weka classification

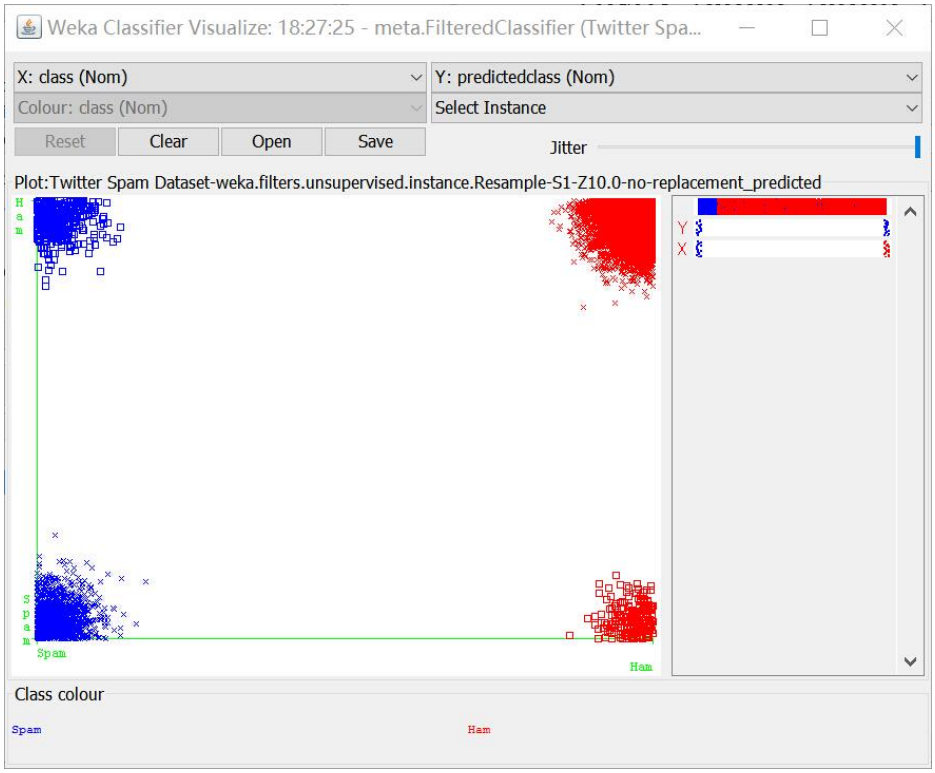

Figure 3-5. Classification Errors of J48 Tree via Weka

Table 3.1 shows the performance of five classifiers. We can see that most of them show reasonable performances with accuracies above 90% but lower recalls. That means these classifiers gained better performance in the ham tweets but mediocre performance in the spam tweets. Among them, tree classifiers (Random Forest and Random Tree) show better performance than others. Random Forest obtains the best performance with nearly 95% in accuracy, precision as 98.5%, and 0.66 in Recall. Logistic shows the best performance in spam detection with it owns the highest recall (0.68), but its final accuracy (94.1%) is a bit lower than Random Forest. BayesNet gains a nearly 90.6% in accuracy, 74.2% in precision, and the lowest recall among the five classifiers which nearly 0.582. Naïve Bayes gains 90.4 % in accuracy, 70.9% in precision, 0.615 in recall and 0.659 in FM. Table 3.2 shows the influence of training set's size to the final classification results. When 20% data is used for training, the final accuracy is up to 0.962, and the false positive is as low as 0.208. Only a small part of spam achieves high accuracy in the entire spam detection.

Table 3.1. Performance results of TSD

Classifier	**Accuracy**	**Precision**	**Recall**	**FM**
Random Forest	0.947	0.985	0.66	0.79
Logistic	0.941	0.908	0.680	0.778
Random Tree	0.927	0.808	0.675	0.735
BayesNet	0.906	0.742	0.582	0.652
Naïve Bayes	0.904	0.709	0.615	0.659

Table 3.2. Classification of TSD via Random Forest

Percentage	Accuracy	Precision	Recall	FM
5%	0.945	0.891	0.729	0.802
10%	0.947	0.985	0.66	0.79
15%	0.956	0.984	0.721	0.832
20%	0.962	0.986	0.755	0.855

3.2 Facebook Original Classifications

Table 3.3 shows the results of FSD which uses several classifiers for detection. The results show these classifiers obtain accuracies all higher than 90%, but recalls are all lower than 90%. The tree classifiers (Random Forest, J48) show better performances compared to others. Random Forest shows the best performance with the accuracy close to 0.977, a nearly 0.844 recall and precision is high as 0.928. It can also show that Bagging shows the best performance in recall which amounts to 0.875, but it shows a lower performance in precision and accuracy. In Table 3.4, it introduces the performance of how training dataset influences the test performance. As the size of train dataset increases, the total accuracy is growing. When 20% of the data was used for training process, the final accuracy reached 98.4%.

Table 3.3. Performance results of FSD

Classifier	Accuracy	Precision	Recall	FM
Random Forest	0.977	0.928	0.844	0.884
Bagging	0.967	0.822	0.875	0.848
J48	0.96	0.793	0.828	0.810
Random Tree	0.949	0.760	0.745	0.753
Logistic	0.924	0.592	0.870	0.705

Table 3.4. Classification of FSD via Random Forest

Percentage	Accuracy	Precision	Recall	FM
5%	0.967	0.862	0.816	0.838
10%	0.977	0.928	0.844	0.884
15%	0.981	0.944	0.868	0.904
20%	0.984	0.951	0.893	0.921

3.3 Classifications across Different OSNs

The next research step was to use the TSD and FSD to continue the mixed classifications. Throughout this part, we modified the detection methodology. We combined the original datasets with the spams from other OSNs. When we split the TSD into training and test sets, we combined spam of Facebook with the Twitter dataset as Twitter Spam Mixed Dataset (TSMD), and then we used it to train classifiers and do the detections. Also, this research got the Facebook Spam Mixed Dataset (FSMD) where we combined spam of Twitter with the Facebook dataset. For classifying spam and ham compared with prior

detection, we recalculated the probabilities for all contents. Now we introduce the parameter RP_i (Revised Probability), defined as follows:

$$RP_i = \alpha P_{1i} + \beta P_{2i}$$

In the above formula, P_{1i} is the probability of i^{th} content during the classification of original data sets, P_{2i} is the probability of i^{th} content via the classification combined with outside spam. α and β are ratios for the spam or ham. This research mainly focuses on the result of spam classification of new datasets, so when calculating the probabilities of new spam which was marked as ham by original classifications, the progress will set α and β as 0.8 and 0.2, because specify ham mainly based on the original datasets.

From Table 3.5, after combining with spam of Facebook; the final precision is nearly 1.4% to 2.6% better than before. For the recall, all five classifiers perform more than 15% better than before. Naïve Bayes acquires the most increase in the recall - up to19.8%. Random Forest shows the best overall performance and has accuracy up to 97.3% and nearly 16.1% increase in the recall. Figure 3-6 shows the difference of false classified spam, which means, those spam posts had been classified as ham. For Random Forest, the final number of false classified spam has declined from 601 to 297, achieving a 51% reduce. The spam of Logistic declines nearly 54.8% from 560 to 253. The spam of Random Tree and Logistic separately reduce by 53.8% (569 to 263) and 54.8% (560 to 253). Also for Naïve Bayes and BayesNet, they have reduced 25.4% (from 674 to 503) and 35.2% (733 to 475).

Table 3.6 shows the result of FSMD. It shows that when combined with spam of Twitter, all classifiers show increases in the accuracy and recall in various degree. Random Forest shows the best performance with its accuracy is up to 98.9%, and recall is nearly 85.5%. It

gains an 11.1% increase in recall compared with the original classification. Logistic gains a most rise in recall which is up to 16.3%. Random Tree and Bagging have 1.4%, 1.2% increases in accuracy and 13.1%, 9.6% rise in the recall. J48 obtains a 0.3% growth in accuracy and 3.4% growth in the recall. Figure 3-7 shows the final numbers of false classified spam in FSD and FSMD. Random Forest has a nearly 71.2% reduce in spam from 146 to 42. Random Tree, Logistic, and J48 has 51.3%, 49.2% and 47.2% reduce in spam from 238 to 116, 122 to 62 and 161 to 85, respectively. Bagging has nearly 39.3% reduce in classified spam from 117 to 71.

Table 3.5. Classification of TSMD via Random Forest

Classifiers	**Results of TSMD**		**Results of TSD**	
	Accuracy	***Recall***	***Accuracy***	***Recall***
Random Forest	0.973(2.6%↑)	0.831(16.1%↑)	0.947	0.66
Logistic	0.967(2.6%↑)	0.855(17.5%↑)	0.941	0.680
Random Tree	0.953(2.6%↑)	0.850(17.5%↑)	0.927	0.675
BayesNet	0.928(2.2%↑)	0.729(15.7%↑)	0.906	0.582
Naïve Bayes	0.918(1.4%↑)	0.713(19.8%↑)	0.904	0.615

Table 3.6. Classification of FSMD via Random Forest

Classifiers	**Results of FSMD**		**Results of FSD**	
	Accuracy	***Recall***	***Accuracy***	***Recall***
Random Forest	0.989(1.2%↑)	0.955(11.1%↑)	0.977	0.844
Bagging	0.972(1.2%↑)	0.924 (9.6%↑)	0.96	0.828
J48	0.970(0.3%↑)	0.909(3.4%↑)	0.967	0.875
Random Tree	0.963(1.4%↑)	0.876(13.1%↑)	0.949	0.745
Logistic	0.931(0.7%↑)	0.933(16.3%↑)	0.924	0.870

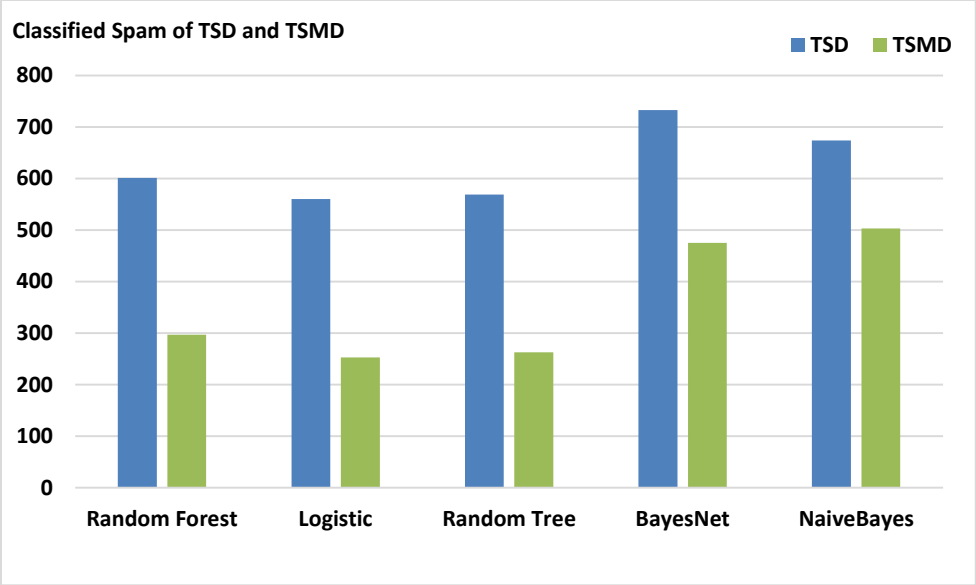

Figure 3-6. Number of false classified spam in TSD and TSMD

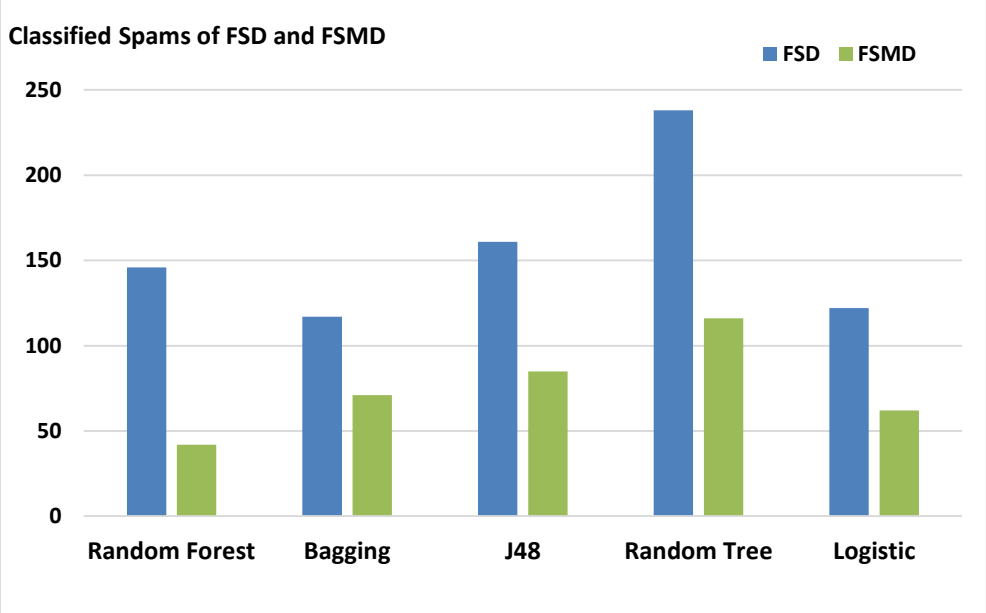

Figure 3-7. Number of false classified spam in FSD and FSMD

Figure 3-8 shows the number of false classified Twitter spam in TSD via classifications combining with different sizes of Facebook spam. False classified means those spam posts

had been classified as ham. All five classifiers show significant declines in the final number of spam. The total number of spam via Random Forest declines from 601 to 282, achieve nearly 53.1% reduce. The total number of spam via Random Tree declines from 569 to 263, achieve nearly 53.8% reduces. The total number of spam via Logistic declines from 560 to 253, achieve nearly 54.8% reduces. The total number of spam via BayesNet declines from 733 to 475, achieving nearly 35.2% reduces. Naïve Bayes achieves a decline of spam from 674 to 503, which has nearly 25.4% reduces. Random Forest, Random Tree, and Logistic obtain better performances for classifying spam than Naïve Bayes and BayesNet.

Figure 3-9 shows the number of false classified Facebook spam posts in FSD when combining with various size of Twitter spam. These classifiers all show a decrease in the number of spam posts, while Random Forest, Random Tree, Logistic, and J48 show better performance. The total number of spam via Random Forest declines from 146 to 42, achieve nearly 71.2% reduces. The total number of spam via Random Tree declines from 238 to 116, achieve nearly 51.3% reduces. J48 achieves a decline of spam from 161 to 85, which has nearly 47.2% reduces. Bagging has a decline of spam from 117 to 74, which has nearly 36.8% reduces. The number of spam for logistic has a decline of spam from 122 to 62, which has nearly 49.2% reduces.

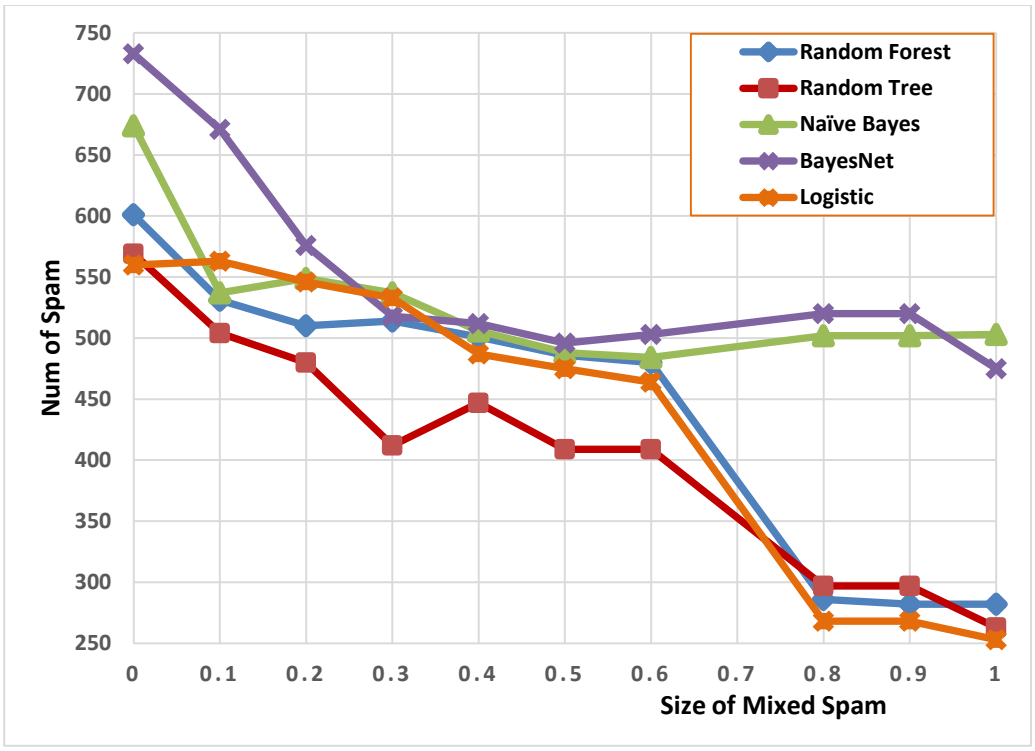

Figure 3-8. Number of false classified spam via different size of spam from Facebook

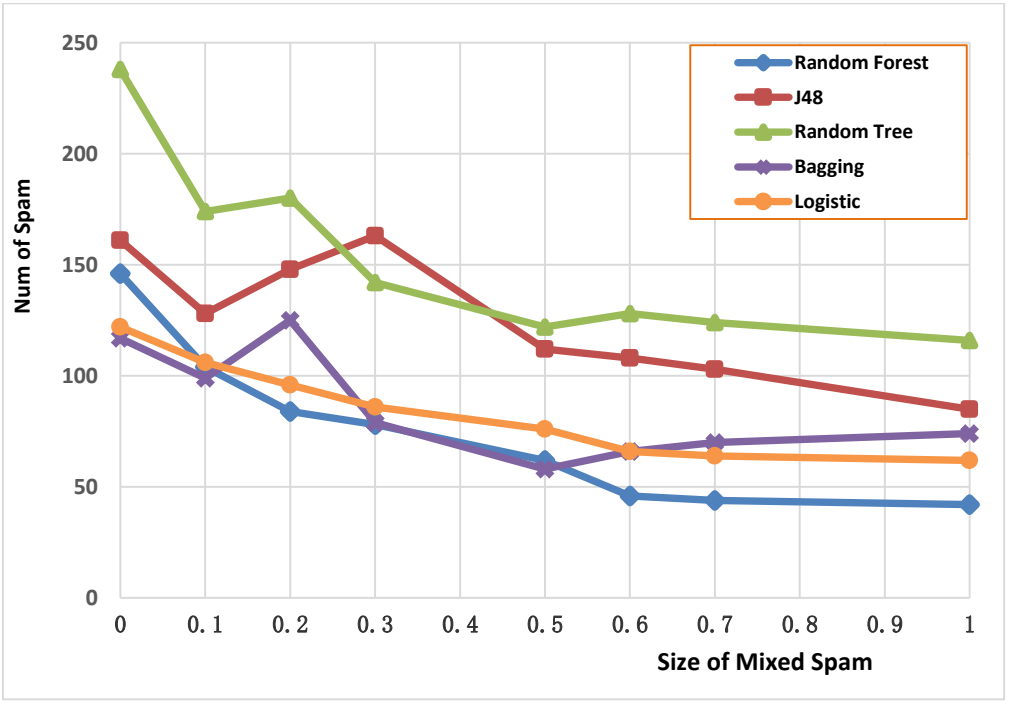

Figure 3-9. Number of false classified spam via different size of spam from Twitter

Chapter 4

Conclusions

4.1 Introduction

Online social networks spam detection and classification have been a popular topic in the science and technology areas. Scientists and researchers pay lots of attention in it to build a more developed and convenient visual world to human beings. This research makes a new progress in this topic and proposes a new point of spam detection.

4.2 Conclusion of Research Performance

In this whole research, we introduce a new perspective to distinguish between spam and legitimate contents in Twitter and Facebook, the top two most popular social networks in the world. For research convenience, we collected two new datasets through their APIs via similar topics and users group. When collected Twitter dataset which the keyword was set as “Taylor Swift” on Twitter from June 2015 to August 2015, and then get the Twitter Spam Dataset (**TSD**). After our labeling and normalizing, we got this dataset that consists of 1937 spam tweets and 10942 ham tweets. For Facebook, We collected data from the open public group on Facebook, which was named as “World of Taylor Swift” from July

2015 to August 2015, and then get the Facebook Spam Dataset (**FSD**). We labeled and normalized them into one dataset that contains 1338 spam posts and 9285 ham posts.

The whole classification progress through via Weka [13], a tool contains several traditional classifiers such as Random Forest, Random Tree, J48, Logistic, and Naïve Bayes, to evaluate these two original datasets. Those classifiers achieve reasonable and well performance in major features like Precision of spam, F1-Measure, Recall, and accuracy.

Random Forest shows the best performance in accuracy and recall for Twitter Spam dataset, also the best performance in accuracy and recall for Facebook Spam dataset. We then combined the spam of one social network with another to enhance the training work for classifications. As former design expected, the spam related with same topics in another OSN lead a positive influence in the spam detection in this OSN. The new classifications with mixed spams from another social network show better performances in both precision and false positive. The results show that Random Forest obtains the best performance with the accuracy and recall for the TSD and FSD, and achieve the most decrease in the number of spam detected. The results demonstrated that similar spam in one online social network benefits the spam detection in another social network. Our future research will focus on analyzing more inner-connections and activities of spammers in different OSNs. We intend to build a multi-function framework that could efficient classify various types of spam across different OSNs.

References

[1] Wikipedia. ‘Social network service’, 2016. [Online]. Available: https://en.wikipedia.org/wiki/Social_networking_service. [Accessed: 27-July-2016]

[2] B. Logan, ‘Publishing to Twitter from Facebook Pages’, 2009. [Online]. Available: https://www.facebook.com/notes/facebook/publishing-to-twitter-from-facebook pages/123006872130. [Accessed: 27-July-2016]

[3] Neal Ungerleider, ‘Almost 10% of Twitter is spam’, 2014. [Online]. Available: http://www.fastcompany.com/3044485/almost-10-of-twitter-is-spam. [Accessed: 27-July-2016]

[4] Avoiding Social Spam Hackers on Facebook and Twitter. [Online]. Available: http://www.sileo.com/social-spam/. [Accessed: 27-July-2016]

[5] C. Nerney, ‘5 top social media security threats’, 2011. [Online]. Available: http://www.networkworld.com/article/2177520/collaboration-social/5-top-social-media-security-threats.html. [Accessed: 27-July-2016]

[6] X. Hu, J. Tang, and H. Liu. “Online social spammer detection,” In 28th AAAI Conference on Artificial Intelligence, 2014.

[7] X. Jin, C. Lin, J. Luo, and J. Han. “A data mining-based spam detection system for social media networks,” In Proceedings of the VLDB Endowment, 4(12), 2011.

[8] B. Markines, C. Cattuto, and F. Menczer. “Social spam detection,” In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, 2009, pp. 41-48.

[9] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, 2010, pp. 35-47.

[10] K. Thomas, F. Li, C. Grier, and V. Paxson. "Consequences of connectivity: Characterizing account hijacking on Twitter," In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp. 489-500.

[11] C. Grier, K. Thomas, V. Paxson, and M. Zhang. "@ spam: the underground on 140 characters or less," In Proceedings of the 17th ACM conference on Computer and communications security, 2010, pp. 27-37.

[12] D. Wang, D. Irani, and C. Pu. "SPADE: a social-spam analytics and detection framework," Social Network Analysis and Mining, 4(1), 2014, pp. 1-18.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, 11(1), 2009, pp. 10-18.

[14] M. Bosma, E. Meij, and W. Weerkamp. "A framework for unsupervised spam detection in social networking sites," In Advances in Information Retrieval, 2012, pp. 364-375.

[15] J. Caverlee, L. Liu, and S. Webb. "Socialtrust: tamper-resilient trust establishment in online communities," In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, 2008, pp.104-114.

[16] J. Song, S. Lee, and J. Kim. "Spam filtering in Twitter using sender-receiver relationship," In Recent Advances in Intrusion Detection, 2011, pp. 301-317.

[17] K. Thomas, C. Grier, D. Song, and V. Paxson. “Suspended accounts in retrospect: an analysis of Twitter spam,” In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, 2011, pp. 243-258.

[18] K. Lee, J. Caverlee, and S. Webb. “Uncovering social spammers: social honeypots+ machine learning,” In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 435-442.

[19] Y. Zhu, X. Wang, E. Zhong, N. N. Liu, H. Li, and Q. Yang. “Discovering Spammers in Social Networks,” In AAAI. Dec 2012.

[20] S. Long, R. Y. Lau, and C. Yin. “Discriminative topic mining for social spam detection,” In PACIS, 2014, pp, 378.

[21] J. Bowden, ‘Social Media APIs and Data Collection Strategies’, 2014. [Online]. Available: http://www.business2community.com/social-media/social-media-apis-data-collection-strategies-0887426#DSb12I7Hbs7XTAOr.99. [Accessed: 27-July-2016]

[22] Internet Social Networking Risks. [Online]. Available: https://www.fbi.gov/about-us/investigate/counterintelligence/internet-social-networking-risks. [Accessed: 20-May-2016]

[23] Harris Interactive Public Relations Research. “A study of social networks scams,”. 2008.

[24] Twitter Company. ‘Twitter API Overview’, 2016. [Online]. Available: https://dev.twitter.com/overview/api. [Accessed: 27-July-2016]

[25] Facebook Company. ‘Facebook for developers’, 2016. [Online]. Available: https://developers.facebook.com/. [Accessed: 27-July-2016]

[26] G. Stringhini, C. Kruegel, G. Vigna. “Detecting spammers on social networks,” Proceedings of the 26th Annual Computer Security Applications Conference. ACM, 2010, pp, 1-9.

[27] Wikipedia. ‘Sensitivity and specificity’. [Online]. Available: https://en.wikipedia.org/wiki/Sensitivity_and_specificity. [Accessed: 27-July-2016]

[28] J. Pei, B. Zhou, Z. Tang, D. Huang. “Data Mining Techniques for Web Spam Detection,” . Simon Fras University Microsoft Ad Center.

[29] Wikipedia. ‘Naive Bayes classifier’. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Constructing_a_classifier_from_the_probability_model. [Accessed: 27-July-2016]

[30] Wikipedia. ‘C4.5 algorithm’. [Online]. Available: https://en.wikipedia.org/wiki/C4.5_algorithm. [Accessed: 27-July-2016]

[31] Wikipedia. ‘Random tree’. [Online]. Available: https://en.wikipedia.org/wiki/Random_tree. [Accessed: 27-July-2016]

[32] Wikipedia. ‘Random forest’. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest33. [Accessed: 27-July-2016]

[33] Wikipedia. ‘Bayesian Network’. [Online]. Available: https://en.wikipedia.org/wiki/Bayesian_network. [Accessed: 27-July-2016]

[34] Wikipedia. ‘Logistic Regression’. [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression. [Accessed: 27-July-2016]

[35] Wikipedia. ‘Anti-spam techniques’. [Online]. Available: https://en.wikipedia.org/wiki/Antispam_techniques#Detecting_spam. [Accessed: 27-July-2016]