MATCHING OF SOCIAL MEDIA ACCOUNTS BY USING PUBLIC INFORMATION

A Thesis Submitted to the Graduate School of Engineering and Sciences of İzmir Institute of Technology in Partial Fulfillment of the Requirements for the Degree of
MASTER OF SCIENCE
in Computer Engineering

by YAĞIZ ÇETİNKAL

October 2016 İZMİR

We approve the thesis of **Yağız ÇETİNKAL**

Examining Committee Members:

Asst. Prof. Dr. Serap ŞAHİN
Department of Computer Engineering, Izmir Institute of Technology

Asst. Prof. Dr. Mutlu BEYAZIT
Department of Computer Engineering, Yaşar University

Asst. Prof. Dr. Belgin Ergenç BOSTANOĞLU
Department of Computer Engineering, Izmir Institute of Technology

19 October 2016

Asst. Prof. Dr. Serap ŞAHİN
Supervisor, Department of Computer
Engineering, Izmir Institute of Technology

Assoc. Prof. Dr. Y. Murat ERTEN
Head of the Department of Computer Engineering

Prof. Dr. Bilge KARAÇALI
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Asst. Prof. Dr. Serap Şahin for her guidance throughout the time of preparation of my thesis. I thank her for ultimate patience and encouragement which makes this thesis possible to be finished.

I also thank Asst. Prof. Dr. Selma Tekir for her support and guidance

.

ABSTRACT

MATCHING OF SOCIAL MEDIA ACCOUNTS BY USING PUBLIC INFORMATION

Protection of private information on social networks (SNs) has become a serious and important topic since social network sites became popular and widely adopted worldwide. Usually people want their personal information to be known only by a small group of people including close friends and families. But sometimes they willingly accept to give some particular information about themselves to individuals which are neither a friend nor an acquaintance. Each SN has different purposes and people subscribe many of them. However, public information available on these sites reveals many aspects of user’s identity. In this work, it is shown that public information can be used to detect the different accounts of the same individual.

This study is performed on two popular social media sites: Twitter and Facebook. Public attributes of the profiles such as real name, user name and status updates (tweets and posts) are used for comparing profiles on two SNs. Different data mining algorithms are compared for matching profiles. Also relationship between text similarity and total term counts of status updates is analyzed.

Results show that simple features like real names, user names and status updates have high similarity between the accounts of the same users and these features can be used to detect profiles of the same user on different SNs. Also the more status updates a user posts on Facebook the more he will likely be detected by the matching schema. Thus, public information can be exploited to pose a threat to the privacy of the people on the Internet.

ÖZET

SOSYAL MEDYA HESAPLARININ HERKESE AÇIK BİLGİLERİN KULLANILARAK EŞLEŞTİRİLMESİ

Sosyal ağlar dünyada popüler ve yaygın olduğundan beri gizliliğin korunması ciddi ve önemli bir konu olmuştur. Genellikle insanlar kişisel bilgilerini sadece yakın arkadaşların ve ailelerinin dâhil olduğu küçük bir grup ile paylaşır. Fakat bazen kendileri hakkındaki bazı bilgileri isteyerek yabancılarla da paylaşmak isteyebilirler. İnsanlar farklı kullanım amaçları olan birçok sosyal ağa kaydolmaktadır. Fakat sosyal ağlardaki herkese açık olan bu bilgiler kullanıcıların kimliğinin birçok noktasını açığa çıkarmaktadır. Bu çalışmada, herkese açık bu bilgiler kullanılarak aynı kişinin farklı sosyal ağ hesaplarının keşfedilebilir olduğu gösterilmektedir.

Çalışma en popüler sosyal ağlardan Twitter ve Facebook üzerinde gerçekleştirildi. Hesaplardaki gerçek isim, kullanıcı ismi ve durum güncellemesi (tweetler ve yazılar) gibi herkese açık bilgiler, iki sosyal ağ üstündeki hesapların karşılaştırılması için kullanıldı. Hesapları eşleştirmek için farklı veri madenciliği algoritmaları karşılaştırıldı. Ayrıca hesaplar arasındaki yazı benzerliği ile yazılardaki terim sayısı arasındaki ilişki incelendi.

Sonuçlar, aynı kişinin farklı hesapları arasında gerçek isim, kullanıcı ismi ve durum güncellemesi gibi basit niteliklerin yüksek oranda benzerlik gösterdiğini ve bu niteliklerin aynı kişilerin farklı sosyal ağlardaki hesaplarını tespit etmede kullanılabileceğini göstermektedir. Ayrıca kullanıcılar Facebook'da ne kadar çok yazarsa, Twitter hesabı ile eşleşme olasılığı o kadar artmaktadır. Sonuç olarak herkes tarafından erişilebilen bu bilgiler internetteki kullanıcıların gizliliğine tehdit oluşturacak şekilde istismar edilebilir.

TABLE OF CONTENTS

APPENDICES

LIST OF FIGURES

LIST OF TABLES

This page is left blank intentionally.

CHAPTER 1

INTRODUCTION

Protection of private information on social networks (SNs) has become a serious and important topic since social network sites became popular and widely adopted worldwide. Usually people want their personal information to be known only by a small group of people including close friends and families. But sometimes they willingly accept to give some particular information about themselves to unknown individuals. This information can be very useful to meet new people, to discuss ideas, to learn new things and to expand their network. Also information revealed by users and interactions between them are very useful for companies for marketing purposes and data mining. However this situation raises serious concerns about the privacy of the users. This is directly related to the matching anonymous user profiles from different SNs. In this thesis, privacy risks on social networks, personal identifiable information and data matching challenges are evaluated. Aim of this work is that, it is possible to match user accounts of the same individuals across different social networks using data matching techniques.

In this chapter; definition of SNs, privacy settings on SNs, data stored about users and privacy concerns about user data are discussed. It is followed by the fact that, even user data is considered anonymous, with the help of different data sources, private information about users can be identified. It is shown that account matching problem is another instance of the Entity Resolution and challenges that make the problem harder are discussed. In Chapter 2, related works about matching accounts on SNs are evaluated. In chapter 3, how dataset is constructed, which attributes and SNs are used for analysis is stated. In Chapter 4, methodology and the contributions of the study is explained. In Chapter 5, features used for profile similarity and similarity metrics are stated. In Chapter 6, results of matching are evaluated and improvements are proposed to increase matching rate. In Chapter 7, results of the study are evaluated and recommendations for users to avoid detection from matching are given. In Chapter 8, future works to improve the performance of the matching scheme is discussed.

1.1.Social Networks

Social networks are web based services that allow individuals to create a public or semi-public profile, manage a list of profiles with whom they share a connection, view and traverse others list of connections. Each site has its own nature and different titles for their features. Since their introduction, they have been used by billions of users around the world and many of the users have integrated these sites into their daily practices. While some sites help users to keep in touch with their pre-existing social network, some of them help other members to connect with each other based on their shared interests, political views and activities [1] .

When a user joins a SN, he is asked to answer a series of questions. These questions are typically about name, username, birth date, location and interest areas of the user. According to the answers given by the user a profile is created then the user is usually directed to upload a profile photo and find their connections on the site. These connections may be labeled different for each network and popular ones are friends, contacts and followers. Some SNs require bidirectional confirmation for friendship but some do not. Unidirectional connections are usually called as fans or followers.

The visibility of a profile changes according to the user preferences and default privacy settings of the site. Some sites offer users to hide their profiles or specific parts of profiles from specific users, networks and search engines. Users can hide their whole profile from other users and search engines which make the users totally unreachable. Alternatively user can make his profile public but parts of his profile (such as birthdate, list of friends, photos etc.) visible only to friends. But these privacy settings depend on the site and each of them has different options that can be offered to user. For example, Twitter asks users if they want their profile to be public or private. If a profile is public, all tweets (status updates), responses, list of followers of the profile, list of profiles followed by the account (following list), username, real name and profile photo becomes available for all Twitter users and internet. If a profile is private, only allowed followers of the profile can see the tweets, the list of followers and the followings (profiles followed by the user) of the user. Other features of the profile are still visible to other users such as name, username and profile photo. On the other hand, LinkedIn offers users to create two profiles: one profile for public view where search engines, non-members and non-connected profiles can see and one profile only for connected-profiles which requires bidirectional confirmation. Facebook offers more detailed privacy settings for the users. Visibility of each part of the profile is determined by the user. User can make a part of profile visible to public (this group covers non-members of Facebook and non-friends of the user), friends, friends of friends, a specific network or a sublist of friends.

1.2.Privacy Issues

As more people start to use social networks and number of social network services increase, it raises many privacy concerns. Some of these privacy concerns are about measurement of privacy, control over shared content and sharing of personal data with third parties.

Since the internet became popular and people started to register online websites, their personal information including email addresses, name and address data are stored in servers which are turned into a great potential opportunity for marketers. Usually registered users of websites are asked to fill in a complete profile containing private information. Websites claim that the requirement of filling a complete profile is needed in order to improve their service however websites can use and distribute this data to third parties for marketing and advertising purposes [2].

Majority of the internet economy relies on online advertising. Targeted advertisement ensures advertisers to reach correct consumers and minimizes wasted advertising costs. However it requires collection of large amounts of personal data of internet users which leads to loss of privacy. This situation is summed up by famous phrase “if you are not the consumer, then you are the product” or “if you are not paying the product, you are the product” which tells personal data is traded for free services or products [3] [4].

Most online services require user’s personal information to give service however users are not helpless to protect their privacy against greedy organizations that seeking more personal data. Since media coverage increases about the potential threats about security and privacy on the internet, users started to provide incomplete information to web sites and they are less likely to register for websites requesting information. Some organizations provide notice to users about their information practices and privacy policy. Also mandatory government rules may dictate companies how to collect and use information. Other kind of threat to user privacy is direct attacks to websites which results in stolen user data and malicious applications downloaded by the users [2].

1.3.Personally Identifiable Information

Personal identifiable information (PII) is one of the most widely used term to describe information about a person. Some common examples are name, citizenship number and email address. According to National Institute of Standards and Technology (NIST) definition of PII is "any information about an individual maintained by an agency, including any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information" [5].

"Linked information" is the information about an individual that is logically associated with other information about the individual. In contrast, "linkable information" is the information about an individual where there is a possibility of logical association with other information about the individual [5] .

Two PII elements that belong to the same individual are considered as linked if they are present on the same system or a closely-related system or if there is no effective security controls that can separate the information sources. They are considered as linkable if they are stored remotely, unrelated or one of them is publicly accessible.

Many websites and SNs use third-party servers to provide content and advertisements for users of the first-party servers. Some third-party servers are aggregators which track and aggregate user viewing and browsing habits across multiple sites with the help of tracking cookies. Many popular sites use third-party tracking services and some trackers dominate across a number of popular SN sites [6]. If a user visits at least one SN and reveals a few PII of him, third parties would be able to associate the habits of the user with a specific person [7]. There is no guarantee these data cannot be sold to other third party entities and used for activities such as identity theft, social engineering attacks, online and physical stalking. Some of the leaked information of the users during their online activities to third parties are listed below [8]:

- User agent: It can be extracted from HTTP request and gives information about browser type and operating system information.

- IP Address: This information can be used to perform Geo-Targeting and help to customize the advertisements according to their geographic location.
- Referrer: It gives information about where user is directed from and can be used to infer user's movements and habits.
- Identifying URLs: Some URLs may contain query strings which include personal information.
- History of visited links and cached objects: Script executions can allow to access to browser information.

1.4.Re-identifying of de-identified data

Sometimes data holders can publish collected data of the individuals after de-identifying it to protect the privacy of the entities in the dataset. Removing explicit identifiers such as name, email address, phone number and address from a dataset might make it seem anonymous however combinations of other attributes can be used to re-identify individuals by the help of another dataset. The National Association of Health Data Organizations in US reported that 44 states have mandatory laws to collect medical data from hospitals, physician offices and clinics [9]. Many of these states distribute copies of the data to researchers, sells to industry and make them publicly available. These data usually cover the patient's ZIP code, birth date, gender, ethnicity and no explicit data. Also public voter registration list dataset covers name, address, ZIP code, birth date and gender information of each voter as seen in Figure 1. Sweeney [9] showed that 87% of the population in US had reported characteristics that likely make them unique based only on three attributes: ZIP code, birth date and gender. By the help of linking voter list dataset with medical dataset, de-identified data in medical dataset is re-identified. If there are more attributes with high discriminability then it is more likely those attributes combine uniquely to identify the entity. For example gender has 2 possible values, five-digit ZIP code has 5 possible values and birth date has 36,500 possible values within the range of 365 days and 100 years. Therefore this data set can uniquely identify 365,000 entities. If there are less than 365,000 entities in the dataset, many entities will have unique attribute combinations. Unlikely any entity will share the same tuple. Also, a single attribute may be powerful due to unusual frequency distribution of the attribute value. Birth years with earlier years will tend to occur less and a person born in 1900 will be unusual and less anonymous. Since attributes such as, name, phone number and email address which can be used to reach a person directly and uniquely is defined as explicit identifier, therefore, attributes that associates uniquely or almost uniquely to an entity is defined as Quasi-identifier [9].

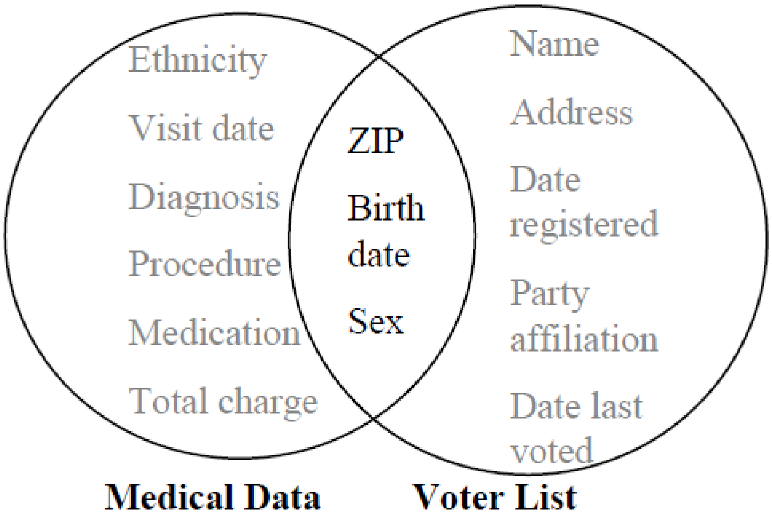

Figure 1. Two datasets used by Sweeney to de-identify individuals [9]

Sweeney proposes a formal protection model named k-anonymity to guarantee that the individuals who are the subjects of the data, can’t be re-identified while the data remain practically useful. If data holder applies k-anonymity model to the data to be shared, each person can’t be distinguished from at least k-1 individuals who also appears in the data. For example if exact birthdate data is replaced with the ranges of years of birthdates, it becomes harder to distinguish individuals from each other in the data set and to re-identify the de-identified data [10].

1.5.Entity Resolution

The task of identifying records that refers to the same real-world entity across different data sources is referred as entity resolution. In many application domains the task has different names such as record linkage, duplicate detection, reference resolution, reference reconciliation, fuzzy match, objection identification, object consolidation, deduplication, object identification, approximate match, entity clustering, identity uncertainty, merge/purge, household matching, hardening sort databases and reference matching [11]. Entity resolution can be classified into two types. First one is pair-wise whose the result is a pair of data objects which refer to the same real-world entity. Record linkage, fuzzy match and approximate match often refer to pair-wise entity resolution. Second one is group-wise whose result is a family of clusters with each one containing the data objects referring to the same real-world entity. Duplication detection and hardening sort databases often refer to group-wise entity resolution. An entity usually corresponds to individuals from different domains such as healthcare, commerce where an entity is a patient or customer. Besides individuals, entities can be records about businesses, publications, citations, products and web pages. Even tough entity resolution has been studied for a long time, entity research has following challenges [11] [12]:

- **Lack of Unique Entity Identifiers and Data Quality**: If all records include unique entity identifiers or keys such as citizenship number, tax payer number or product, then problem becomes a database join problem and it can be implemented efficiently through SQL statements. If no unique identifiers are available then we need to rely on the attributes that are common across the databases.
- **Big Data and Computation Complexity**: Since many applications have big data, processing data efficiently and effectively is an important challenge. Each record from one database needs to be compared with all records in the other database. Number of true matches grows linearly with the size of the databases to be matched, on the other hand number of false matches grows quadratic.
- **Lack of Ground-truth and Training Data**: In many applications it is not known if a matching is a correct match since there is no ground-truth data available that specifies if two records correspond to the same entity or not. Without ground-truth data or extra information provided by surveys or asking individuals about correctness of the match, no one can be sure about the correctness of the outcome of the data matching project.
- **Dynamic Data**: Web pages on internet and financial data update frequently. Current techniques have to scan data multiple times. However current techniques do not support entity resolution on frequently updated data.
- **Heterogeneous Data**: Entities may be represented in different forms of data such as structured, semi-structured and unstructured.

- **Evaluating Results**: Evaluation of entity resolution results efficiently and accurately is another challenge.
- **Privacy and Confidentiality:** Personal information is processed and evaluated, therefore privacy and confidentiality must be carefully considered. The analysis of matched data may uncover individuals' some aspects which are not obvious when a single database is analyzed alone. Sometimes de-identifying the data is not enough to protect privacy, data holder or publisher needs to take some measures [9] [10].

Usually, there are three main steps in methods that propose to match entities. The first step is data pre-processing which is the task of transforming the data from different sources into the same format. In the second step, candidate record pairs are compared using a variety of attributes. In the last step each pair is classified as match or non-match.

1.6.Objectives of the Study

Hundreds of millions of people makes their personal information available on various SNs. Each SN has different purpose, content sharing mechanism and privacy setting. Each profile on these sites reveals a part of the real identity of the creator of the profile. Photographs, interests, opinions, locations, activity time patterns, friend lists are some of the parts of the identity [13]. Figure 2 shows whole online identity of an individual and common attributes available in different networks. Different attributes across different social networks are linkable information of the same individual, because all of them are publicly available. If an attacker can correlate these parts of the identity then attacker may reveal the real identity of the user or may access unlinked personal information about the real identity.

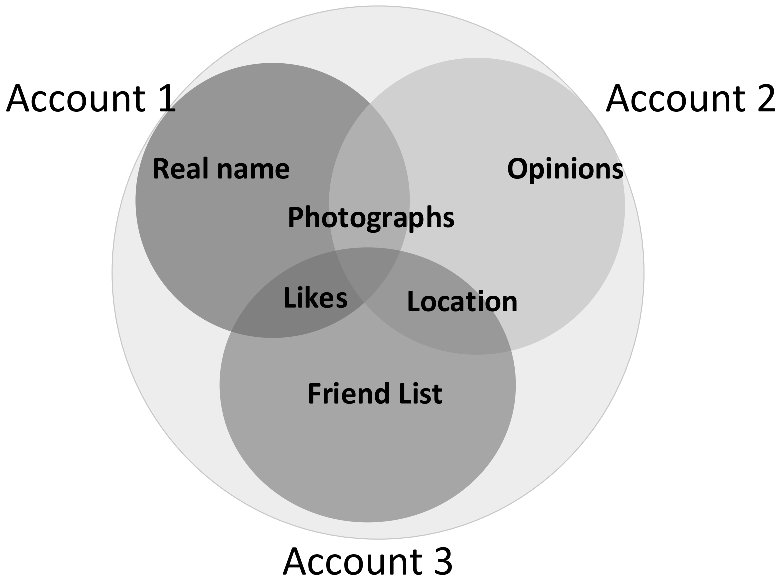

Figure 2. Each social Network account reveals parts of the identity

Seemingly unrelated accounts can be matched using the information provided by the user on different SNs as seen in Figure 3. If accounts are matched, then more personal information of the user can be collected. Therefore, matching accounts of the same person poses a threat to the privacy of the people on the Internet.

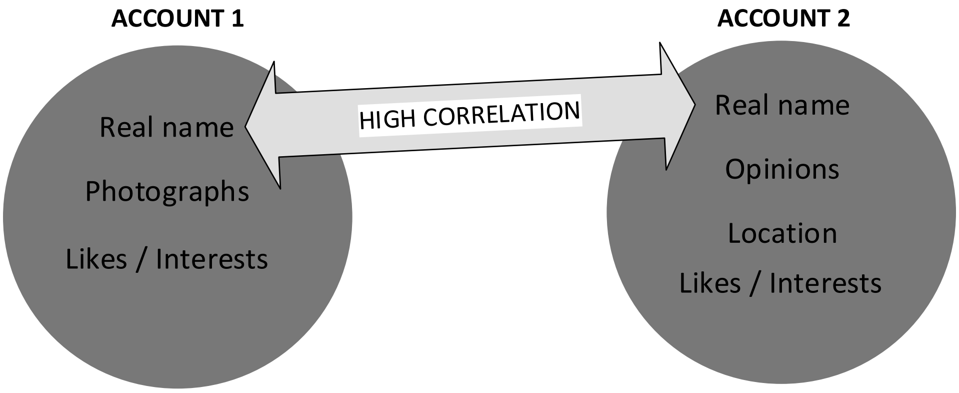

Figure 3. Contents between two social media accounts can be highly correlated

Objectives of the study are to show that there is high potential of identifying independent accounts of the same person in different SNs and to give Internet users recommendations to avoid such matching mechanisms to protect their privacy. Public information which is accessible by search engines, non-friends, non-members and third parties, is available for many SNs and many users. If a user has multiple accounts it is highly possible to match these accounts by comparing appropriate public attributes and with the help of data mining algorithms.

1.7.Challenges of Entity Resolution for Matching Problem in SN

The matching problem is another instance of Entity Resolution. All challenges of Entity Resolution as mentioned in section 1.5 are valid for the matching problem in SNs. Also the additional challenges below make it more difficult to match accounts on SNs:

- **Noisy information:** Same user on different social networks may give different values for the same attributes, or may hide them from public access. Matching scheme requires same or similar attribute values to match the profiles.
- **Large Scale:** A real name attribute can belong to one person in a small dataset however there might be tens or hundreds of people shares the same name in a large scale dataset. That situation yields to increase number of false matches.
- **Data collection constraints:** SNs give limited access to their APIs and resources. APIs may have time limitations to send request or SNs may take precautions to stop crawlers to access and parse profile information.

The reliability and scalability of a matching scheme is highly dependent to the selected features of the accounts. Goga [14] identified four key properties that an ideal feature should have as availability, consistency, discriminability and non-impersonability.

- **Availability:** The selected feature should be available for a large fraction of user accounts. SNs must enable users to share the feature then user must share this feature in both SNs publicly.
- **Consistency:** Users must provide same feature values in both SNs. If they are different then features becomes useless.
- **Discriminability:** A feature with high discriminability would have unique and different value for each user. For example name feature is a more discriminating feature than gender.
- **Non-Impersonability:** A feature shouldn’t be impersonated or faked easily by attackers. Name and profile pictures are some of the features can be copied and used for creating fake accounts.

In this study, networks and attributes are evaluated by their availability, consistency and discriminability because they are directly related with the performance of matching scheme. Impersonators and fake accounts are out of scope for this study because it is assumed that users gave true information about their matching accounts on different SNs.

CHAPTER 2

RELATED WORK

Previous works range between 2012 and 2015 are examined. These studies focus on matching profiles among different social networks using publicly accessible data. Studies differ from each other by data collection methods, selection of social networks, attributes chosen for comparison and evaluation methods of their performances. Performances of matching schemas are evaluated by True Positive (TP) rate, False Positive (FP) rate, recall, precision, Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC). TP rate and recall are ratio of correctly classified instances of the true class to all instances of the true class. FP rate is ratio of instances of false class which are classified as true to all instances of the false class. Precision is the number of TP instances divided by the sum of TP and FP. ROC curves shows tradeoff between TP and FP rates visually and AUC is the total area under the ROC curves.

Almishari et al (2012) study the linkability of reviews authored by the same contributor [14]. Their study is based on over 1,000,000 reviews and 2,000 contributors from Yelp which is a popular review site for local businesses. By the help of Yelp, users can search businesses, give comments about businesses and rate each other's reviews. For each account, reviews are divided into two sets: identified records (IR) and anonymous records (AR). IR for all accounts are used as training set for building models. After matching model is constructed, if an anonymous record is given as input, output is a sorted list of possible accounts. If correct account is on the top of the list with the highest probability then it is considered as perfect hit, if correct account is among the best 10 candidates then it is considered as near hit. If correct account is not in the list then it is a miss. All reviews are tokenized by four types of tokens. They are unigram, digrams, rating and category. Unigrams are set of all single letters and digrams are set of all consecutive letter-pairs. Rating is the score of the review ranges between 1 and 5. Category is related with the service being reviewed. There are 28 categories. Authors also note that they experimented their models on larger token sets like trigrams and stemmed-words (root form of words), however these token sets performed worse than unigrams and digrams. Results showed that unigram tokens obtain high matching scores that reach up to 83% recall and results improve to 96% when rating and category tokens are introduced. They showed that anonymous reviews can be de-anonymized by using simple features. Their results also implicate that accounts of the same person between multiple reviewing sites can be linked since many people tend to maintain their characteristic in writing reviews.

Peled et al (2013) defined the process of identifying different profiles of the same individual as entity resolution [15]. They used dataset consist of 30.000 users collected from Facebook and Xing which is a European social network for business professionals. Totally 27 feature is extracted to compare two profiles. 10 of them are name based features which represent the similarity between two names. There are 15 general user information based features and they represent the similarity between the different parts of personal information of two users. The personal information are extracted from profile pages, which are location, current employer, professional experience and educational background. There are two social network graph based features and they are mutual friends and mutual friends of friends. They calculated two performance score for each future. All-but-x is the score if all features are combined without the feature x and only-x is the score for the feature x alone. Results show that name based features have strong impact on classification since many name based features give highest only-x scores. Classification performance of proposed algorithm measured by AUC was 0.982. Results show that user identification based on web profiles is practically possible.

Soltani et al (2013) classified features of user profiles in three categories as Personal Identities (PI), Social Identities (SI) and Relational Identities (RI) [16]. Personal identities are attributes like name, gender, location, education, email, language and birth date. Social identities are shared contents. Relational identities are friendship graph, group membership and fan page participations. Given an input profile from the source network, a list of candidate profiles are gathered using search operations that are performed by some key attributes using APIs and search engines of destination social networks. Researchers used data set of 20 users for Facebook-Twitter and Facebook-LinkedIn profile matching. Since PI attributes are text based, they compared attributes by exact matching, edit distance and geolocation distance. For comparing SI attributes they used NLP technique to extract topic and category name and YouTube API to extract video category. For comparing RI attributes, number of common friends and membership among profiles are calculated. To make a final decision among candidate profiles they used two approaches. First approach is listing first k similar profiles and the second approach is determining a threshold value and listing profiles with scores that is higher than the threshold value. Their best precision score is about 60% and recall is about 50%.

Researchers also claims their NLP categorization trials shows poor results since online APIs were only able to categorize between 23% and 26% of the shared posts/tweets correctly.

Na et al (2013) classifies profile attributes as decisive and non-decisive [17]. Decisive attributes can identify a user uniquely. If the values are same then they belong to the same person. Non-decisive attributes can be used to make a decision. Decisive attribute can be email and non-decisive attributes can be username and real name. If there is no decisive attribute available in the source user or there is no match in the destination user list by the provided decisive attribute, authors search candidate user set using screen name (username) and full name (real name). There are two datasets used in that research. Profilactic dataset is both ground truth data set and source data set since users can give links of their other social network profiles from their Profilactic profile. Other data set is collected from Flickr. Authors focused on building a linear model and propose a method to adjust attribute weights. Their best recall score is about 85% with 95% precision.

Goga et al (2013) examine more than 200.000 account pairs belong to the same individuals between different social networks [18].They compared user accounts from major social networks Facebook, Twitter, Google+, Flickr and Myspace. Profile attributes used for comparison are username, real name, location and profile photo. They also propose a new feature called cross name which is derived by the measurement of the similarity between the username on one social network and real name on the other social network. They approach the task as a classification problem and train a binary classifier with similarity scores. Google+ dataset is used as ground truth dataset and matching performance of features are compared. Results show that real name feature has the highest matching performance with 80% true positive rate for a 10^{-3} false positive rate. Face similarity feature has the worst performance since face detection algorithm is trained with only one photo for each user and authors believe that it can be more effective if it is trained with more photos. Also location and photo features are not good predictors alone, but if they are combined with other attributes they can improve the performance. By combining all the features they achieved 90% true positive rate. Highest matching performance is between Google+ and Facebook accounts and the lowest performance belongs to comparisons between Myspace and other networks because availability of the real name attributes in Myspace accounts is very low. Authors also show that two accounts which can’t be matched directly in previous research can be matched with the help of a third account from another social network. They use three step correlation chains and match between 6% and 23% of the remaining unmatched account pairs.

Goga et al (2015) propose to evaluate the profile attributes by a set of properties [19]. These properties are availability, consistency, non-impersonability and discriminability as mentioned in section 1.7. A reliable matching schema depends on the attributes that considered for matching and on their properties. Firstly an attribute should be available in both SN and users must provide these information. Matching same person on different networks, highly depends on the consistency of the attribute; it should be same or similar. However sometimes many profiles of different people can share same or similar attributes and a low discriminating attribute can lead to high number of false matches. Also sometimes attackers can create fake accounts of individuals with impersonated attributes such as name and photo which yields to matching of wrong accounts instead of real accounts. Authors also use precision and recall metrics to evaluate the reliability of the matching schemes since true positive and false positive rates are unreliable due to huge class imbalance. If there are 1,000 matching and 999,000 non-matching profiles in a dataset, 90% true positive rate with 1% false positive rate means 900 true matches and 9,990 false matches. As number of profiles increase matching scheme produce more false matches. In a real world application there would be 1 billion non-matching profiles for each profile. While previous studies achieve 90% recall and 95% precision, recall dropped to 19% after schema is applied on full Facebook dataset. Since authors don’t have access to full Facebook database, they exploited the Facebook Graph Search to estimate the discriminability of name features. By proposing a new matching scheme they successfully improved the recall value to 29%. Features are real name, user name, location, profile photo and friends. They tried to find the Facebook profile of a given Twitter profile. For each twitter profile in the sampled dataset, all profiles from Facebook with same or similar real names and user names are retrieved. By doing that they achieved collecting all non-matching profiles with highest similarity and discriminability of the entire social network is preserved.

Previous works are based on small scale datasets that ranges between tens and hundred thousand users. Ground data sets are constructed by the help of links shared by the users willingly on the internet. Relations between profiles of the same individuals in different SNs and all attributes are public. Supervised data mining techniques are used to build a model that matches profiles. Main factors affect the performance of matching schemes are which attributes and how these attributes are used to make a comparison.

Majority of the studies have found that name features are very discriminative and successful than other features. Studies show that it is possible to match user profiles with high recall and precision values on small scale datasets. However in large scale datasets, performance of matching drops dramatically because attributes become less discriminative in a large population. Goga et al (2015) showed 90% recall value drops to 19% after matching schema is applied on full Facebook dataset. Also recall value is more significant to evaluate performance since false positive rate is not significant in large scale datasets. Although introducing new features to matching schema may boost the performance and unmatched accounts can be caught by the new scheme.

CHAPTER 3

DATASETS

SDS (Source network Data Set) and DDS (Destination network Data Set) are required to match accounts across two SNs. Profiles of the first network are represented in SDS and profiles of the second network are represented in DDS. A GDS (Ground truth Data Set) is required to evaluate the performance of the matching process because it shows which accounts belong to same individual. By the help of GDS when two accounts are compared it is already known if it is a correct match or not. Without GDS, a proper model could not be developed and correctness of matches could not be known. After GDS is constructed, profiles at other networks which are used to form SDS and DDS are visited to collect publicly accessible user data.

Some social networks and websites allow users to explicitly list their profiles on the other social networks. One of them is Google+. By knowing that more than 97.000 Google+ profiles are crawled to construct GDS. Crawling started by adding seed accounts to the queue and friends of the visited profiles are added to queue to increase the number of profiles in the dataset. Table 1 shows which social network profiles that Google+ users listed. YouTube is the most listed network with the rate of 58%. Both Google+ and YouTube are products of Google and Google allows users to use YouTube service with their Google+ account. Usually users choose same account to manage two social networks and URLs to both profiles are listed in their profile pages automatically. Twitter profile URL was stated in 19% of accounts and Facebook profile URL was stated in 14% of accounts. 15% of the users listed both their Twitter and YouTube accounts and 12% of the users listed their Facebook and Twitter accounts.

Table 1. Number of social networks of users stated at Google+ profile pages by users

Social Network	**Count**	**Ratio**
Facebook	14165	14.5%
Twitter	18962	19.5%
Flickr	4532	4.6%
YouTube	56809	58.5%
LinkedIn	7735	7.9%
Quora	2135	2.1%
Pinterest	1707	1.7%
Tumblr	1745	1.7%
Twitter and YouTube	15318	15.7%
Facebook and YouTube	11578	11.9%
Facebook and Twitter	11552	11.9%

Users who listed their Facebook and Twitter accounts are chosen for this study because name based attributes and status updates are available in both networks. Attributes available for each network is shown in Table 2.

Table 2. Attributes available for each network

Attributes	**Facebook**	**Twitter**	**YouTube**
Real Name	✓	✓	✓
User Name	✓	✓	✓
Photograph	✓	✓	✓
Status Updates	✓	✓	✗
Location	✗	✓	✗
Gender	✓	✗	✗
Short Description	✗	✓	✓
URL	✗	✓	✓
Time Zone	✗	✓	✗
Locale	✓	✗	✗

To obtain public attributes of Facebook and Twitter users APIs of each network are used. Facebook API give access to default public profile attributes which are always available with no access token [20]. However status updates of the users are not accessible via API even privacy settings allow any user on Facebook to see it, so crawler is used to visit each profile and gather status updates. Profile pictures are gathered by visiting "https://graph.facebook.com/{username} /picture" URL for each username. Attributes available are listed in Table 3.

Table 3. Public attributes of Facebook users

Attribute Name	Description	Type
id	Id of the person's user account	Numeric string
first_name	The person's first name	String
last_name	The person's last name	String
name	The person's full name	String
gender	The person's gender	String
locale	The person's locale. Gives information about user's preferred language. Such as en_US, en_GB, de_DE.	String
username	The person's username. facebook.com/{username} is URL for each profile.	String
updated_time	The last time user updated their profile.	Date time
Status updates	Posts of the person on his wall	List of strings
Profile picture	The person's profile picture	Image

Links on Google + profile pages do not always provide accurate information about the profile of the person on the other SN. In order to evaluate matching of the accounts of the same individual we need a personal Facebook account for each user. There are some reasons why some users are excluded from study even they listed their Facebook accounts. If user listed an invalid URL or listed profile does not belong to an individual

(Facebook Pages and Facebook groups) then it is excluded. These accounts are shown in Table 4. Only 50% listed accounts are valid and belongs to an individual.

Table 4. Number of valid and invalid Facebook accounts in dataset

Reason of exclusion	Count	Ratio
Facebook Pages and Groups	3907	33.8%
Invalid URL	891	7.7%
Valid URL but account is removed or not found	873	7.7%

Twitter API allows us to get public attributes of Twitter users and tweets [21]. If a user have a private profile then only authorized users can follow the user, see his tweets and list of followers. Some attributes available for Twitter users are listed in Table 4. 45% of the listed Twitter accounts are valid. 17% of the accounts are private which means we can’t access their tweets follower lists. Excluded accounts are shown at Table 5.

Table 5. Public attributes of Twitter accounts

Attribute Name	Description	Type
created_at	The UTC date time that the user account was created on Twitter.	String
description	Text describes the account defined by the user.	String
favourites_count	The number of tweets this user has favorited in the account's lifetime.	Integer
followers_count	The number of followers this account currently has.	Integer
friends_count	The number of users this account is following. The other name of this parameter is followings.	Integer
id	Unique user id for Twitter.	Signed 64 bit integer

(cont. on next page)

Table 5. (cont.)

location	User defined location info. It is not necessarily a location.	String
name	Name of the user	String
protected	Indicates if profile is private or not	Boolean
screen_name	Unique username for the account. It can change overtime.	String
status_count	Number of tweets	Integer
time_zone	Time zone that is declared by the user	String
url	A URL provided by the user in association with their profile.	String

Finally we have a GDS with 4273 users that have accessible public Facebook and Twitter profiles. GDS is one to one and onto. That means there is always one true match for each profile in the other set. Both profiles of the users have real name and username attributes. We only have status updates of 849 users due to privacy settings of the Facebook profiles. Common profile attributes are shown in Table 6. There are four attributes available in common however photograph similarity is considered as another subject of research and it is not included in this study.

Table 6. Number of valid and invalid Twitter accounts in dataset

Reason of exclusion	Count	Ratio
Invalid URL	353	3%
Valid URL but account is removed or not found	3954	34.2%
Private Accounts	2066	17.8%

3.1.Data Collection Constraints

APIs of SNs are useful for collecting data however there are some limitations while using them. Twitter API allows data collector to access all attributes of a public profile however Facebook API gives only small portion of publicly accessible data. Even status updates of a profile are public, API requires access permission given by user to data collector (Application owner). There are also time constraints. Twitter allows only 15 requests per 15 minutes. In a single request at most 100 users can be added to query to get profile information and at most 100 tweets can be requested at a time [22] [23].

CHAPTER 4

METHODOLOGY

As mentioned in Chapter 3; there are two sets of profiles from Twitter and Facebook. We have a GDS with 4273 users that have accessible public Facebook and Twitter profiles. GDS is one to one and onto between these source and target sets. That means there is always one true match for each profile in the other set. Twitter is source network and Facebook is destination network. The goal is to match profiles in source network with corresponding profile in destination network. Each profile in source network is compared with each profile in the destination network. The problem is a classification problem where matching of two profiles of the same person is a correct match and matching of others are incorrect match. That means classifier takes two profiles as input and output is a binary class label (true or false). Matching function between source and destination sets is one to one and onto. For n users in source network there are n users in destination network. At the end of classification phase there are n^2 comparisons, n of them are correct matches and the rest are incorrect matches. Table 7 shows number of profiles to be matched and number of comparisons to be classified. As mentioned in Chapter 3, name based attributes are available for 4273 users and all attributes are only available for 849 users. Dataset of 849 users is used for calculating text (status update) similarity

Table 7. Number of profiles and matches used in classification

	Profiles with name based attributes	**Profiles with name based attributes and status updates**
Number of profiles in SDS	4273	849
Number of profiles in DDS	4273	849
Number of pairs to be compared by classifier	18258529	720801
Number of correct matches	4273	849
Number of incorrect matches	18254256	719952

For each source and target profile pair, similarity scores of features are calculated by appropriate similarity metrics which are explained in Chapter 5. Firstly profiles are matched using single feature at a time. Binary classifier decides if a pair is a match according to the threshold similarity score which is determined by a constant FP rate. If score is higher than a threshold value two profiles are labeled as a match, if score is less than the threshold then classifier labels the pair as non-match. If pair belongs to the same individual then it is a true positive and if they don’t belong to the same individual then it is a false positive. Binary classifier shows us how it will perform for each single feature.

After seeing the individual performances of each feature, all features will be combined using different algorithms. Algorithms used are Naïve Bayes, decision tree, logistic regression, nearest neighbor, SVM (Support Vector Machine) and Backpropagation. Ground truth data is split into training and testing sets using 10-fold cross validation.

Performances are evaluated by ROC curves, AUC, TP-FP rates, precision and recall values. Although ROC curves and AUC values are useful to compare different features and different algorithms, recall values for a fixed precision is used to see the real performance of the matching scheme among previous works.

4.1.Contribution of the Study

There are related studies cover different SNs and different attributes. Some of the studies perform analysis on larger scales. Different data mining algorithms are also compared related studies as presented in Chapter 2. In this study:

1. Similarity of status updates as text documents is introduced as a new feature to match profiles.
2. Number of terms (words) posted by users on different SNs and their effect on similarity are analyzed.
3. It is evaluated that how each algorithm finds decision boundaries for classification and which algorithms are the most suitable for matching problem.

4.2.Methods to Control for Threats to Validity

In this section methods about experiment design and precautions for threats to validity are explained. It can help other researchers who wants make similar study and repeat the experiments about SNs. Methods are grouped as data collection, methodology and data processing.

Data collection:

- Randomly chosen seed users are added to dataset and dataset grew by adding other users from friend lists of seed users and their friends.
- English speaking users are included and other languages are excluded from the study.
- It is assumed that users stated correct SN profile links (which belongs to them) on their Google+ page. Effect of impersonated and fake profiles are ignored. If links are broken, invalid or belong to Facebook page they are excluded from the study. If a user states another individual's profile link, matching schema is not capable to detect it.
- Due to privacy policies of SNs it is not possible to share URLs or other attributes of the users which are subjects to the study. However a GDS can be constructed by following the methods explained in Chapter 3. Researchers must be aware data collection can be restricted and permission from data holder can be required.

Methodology:

- 10 fold cross validation is used to train and test the model.
- Because of limited computational resources and quadratic growth in the matching pairs, random sampling is done over incorrect matches for some algorithms. Sample sizes are stated and results on full datasets for other algorithms are also included.

Data processing:

- When calculating name similarity, accents and special language characters are considered as different characters. (c and ç are different characters)
- When calculating text similarity, only characters with UTF-8 encoding is included. Characters such as emoticons, Chinese words are excluded. URLs are accepted as valid terms and included.
- Libraries mentioned in Section 5.3 are used to calculate similarity scores.

CHAPTER 5

ANALYSIS OF FEATURES

To match profiles between source and destination networks, publicly accessible information provided by the users such as real name, user name and status updates are used. User name and real name attributes are used by related studies however status updates are not used to match two profiles between two SNs. In this study similarity between text documents derived from status updates is added as a new feature to improve performance of the matching scheme.

There are three attributes available from source and destination networks. They are real names, user names and text (tweets and status updates). Five features are derived from these attributes to find similarity between two profiles. They are as follows:

- Name based features: RN (real name), UN (user name), CN1 (crossname1), CN2 (crossname2)
- Post based feature: TEXT

RN, UN, CN1 and CN2 are name based features and they are calculated by string edit distance. RN is distance between two real names and UN is distance between user names. Cross name is similarity between user name and real name between different SNs. CN1 is distance between source (Twitter) user name and target (Facebook) real name, CN2 is distance between source (Twitter) real name and target (Facebook) user name. TEXT feature is cosine similarity between source document and target document. We chose those features because name attributes have high availability and consistency among SNs. Facebook and Twitter users will be studied because status updates are only available for these networks.

5.1.Name Similarity

Jaro string distance is a measure of similarity between two strings. It is popular in the area of entity resolution. The distance is normalized and ranges from 0 to 1 while 1 indicates an exact match between two strings. Jaro string distance is used to measure the similarity of names in this work because previous works show that it is the most suitable metric to measure the similarity between names in online applications [24] [25].

For two strings s and t; s' is the characters in s that are common with t; and t' is the characters in t that are common with s. A character is common between two strings if distance between two appearances of the character is less than the length of the shortest string. $T_{s',t'}$ measures the number of transpositions of characters in s' relative to t'. Therefore Jaro distance for strings s and t is calculated by the equation (5.1).

$$Jaro(s,t) = \frac{1}{3}.\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'|-T_{s',t'}}{2|s'|}\right) \quad (5.1)$$

Table 8 shows similarity scores for some string pairs using Jaro string distance. It gives high similarity scores if user names are derived from real names and small differences exists between two names.

Table 8. Jaro Distance scores for some string pairs

String 1	**String 2**	**Jaro Distance**
Yagizcetin	Yagizcetinkal	0.92
Yagizcetin	cetinYagiz	0.87
Yagizcetinkal	Yagiz	0.79

5.2.Text Similarity

Cosine similarity and TF-IDF (term frequency-inverse document frequency) are used to measure the similarity of texts. It is widely used in the information retrieval community [25]. TF-IDF scheme depends on common terms between two documents. Each term has a weight which shows that how important is a word to a document that belongs to a corpus. Term frequency shows how often a term (word) appears in a document. Inverse document frequency shows how uniquely a term appears in the documents in the whole corpus. In our case the corpus is set of all terms appears in tweets and status updates. For each document a vector is created which stores the TF-IDF values for each term. TF-IDF values are calculated by the equation (5.2).

$$(TF - IDF)_{i,j} = tf_{i,j}.idf_i \tag{5.2}$$

$tf_{i,j}$: Term frequency of term t_i for document j

idf_i: Inverse document frequency of term t_i in the corpus D.

$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$

$n_{i,j}$: Number of times term t_i appears in document d_j

$\sum_k n_{k,j}$: Number of total terms in document d_j

$idf_i = \frac{|D|}{c_i}$

c_i: Number of documents in the corpus that contain the term t_i

$|D|$: Number of documents in the corpus. It is twice as user number since two documents are created for each user. One for source SN and the other one is for destination SN.

If a term occurs many times in a document then TF value increase however if the term occurs in many documents then IDF value decrease.

As a result we have a vector of TF-IDF values for each document and the size of the vector is number of total terms in the corpus. To reduce computational complexity of the process stop words are removed from documents. The similarity between two documents $d \ and \ d'$ is computed using cosine similarity measure (5.3).

$sim(d, d') = \frac{v.v'}{\|v\|\|v'\|}$ where $v \ and \ v'$ are TF-IDF vectors of $d \ and \ d'$. (5.3)

5.3.Libraries, APIs and Tools

Datasets are stored in relational database and processed using Object Oriented Programming Language. Raw data collected from APIs and web pages are in the form of JSON (JavaScript Object Notation) and HTML. In the pre-processing data is cleaned, parsed, standardized and finally data is stored into relational database.

- HTMLCleaner [26] is used to collect and parse Google+ data.
- Twitter API, 140dev Streaming API [27], Twitter-API-PHP [28] , IIS Express 8.0 and PHP 5.4 are used to collect and parse Twitter data.

- Facebook Graph API and HtmlUnit [29] are used to collect and parse Facebook data.
- NetBeans IDE 7.4 and Java 7.0 are used for pre-processing and analysis.
- MySQL Server Community Edition Version 5.1.72 is used to store relational data.
- RapidMiner 7.0 Community Edition and Weka 3.6 are used for visualization and analysis.
- Lucene 4.10.3 and Simmetrics 1.6.2 used for calculating name and text similarity.

Data collection and analysis is performed on one PC (Windows 8, 16 GB Memory, Intel i7 2.40 GHz processor) and one Server (Windows 7, 36 GB Memory, Intel Xeon 3.60 GHz processor)

CHAPTER 6

RESULTS

As mentioned in Introduction chapter, the results can be overviewed by three criteria which are availability, consistency and discriminability. The availability of the attributes determines the number of dimensions in the classification problem. Since five features are derived, the problem is classification in a 5 dimensional space. In this section, firstly each feature is evaluated separately, then all features are used to match accounts by data mining algorithms. Recall from Chapter 3, if profile in the source network is matched with the profile of the same person on the destination network then it is a correct match otherwise it is an incorrect match.

Table 9 shows mean and median values of similarity scores for each feature. Mean values give hints about consistency and discriminability of the features. High mean values for correct matches indicate that the feature is very consistent and low mean values for incorrect matches indicate that the feature is very discriminative. As the gap between the distributions of correct and incorrect matches increase, more account will be matched correctly. RN is the most consistent feature among name based features because correct matches have the highest similarity score. More than 50% of the accounts use same real names across two SNs. All name based features have similar discrimination since mean of incorrect matches are very close. Quartile values of similarity scores are shown at Appendix B.

Table 9. Mean and median values of similarity scores

Feature	**Correct Matches**		**Incorrect Matches**	
	Mean	**Median**	**Mean**	**Median**
RN	0.861	1	0.371	0.45
UN	0.723	0.852	0.347	0.43
CN1	0.554	0.601	0.334	0.413
CN2	0.630	0.746	0.338	0.419
TEXT	0.126	0.111	0.051	0.049

Histograms from Figure 4 to Figure 8 show distribution of similarity scores for each feature. Blue colors denote correct matches and red colors denote incorrect matches. X axis is for similarity score where 1 is the highest score and 0 is the lowest score. Y axis denotes the number of matches in log scale. Distribution of correct matches (blue colors) shows how consistent a feature is. More consistent features have distributions highly stacked on higher scores. On the other hand distribution of incorrect matches (red colors) shows how discriminative a feature is. More discriminative features have distributions highly stacked on lower scores.

Threshold value determines the number of TPs and FPs, so choosing the optimum threshold value depends on having maximum number of TPs and keeping number of FPs minimum. Threshold values are chosen such as where false positives rate is 10^{-5} which means we have 182 incorrect matches which are labeled as true-match by binary classifier.

Figure 4 shows the distribution of RN similarity scores. Threshold value is 0.87 and TP rate is 71%. That shows majority of the profiles in our dataset has chosen same or very similar real names for both network.

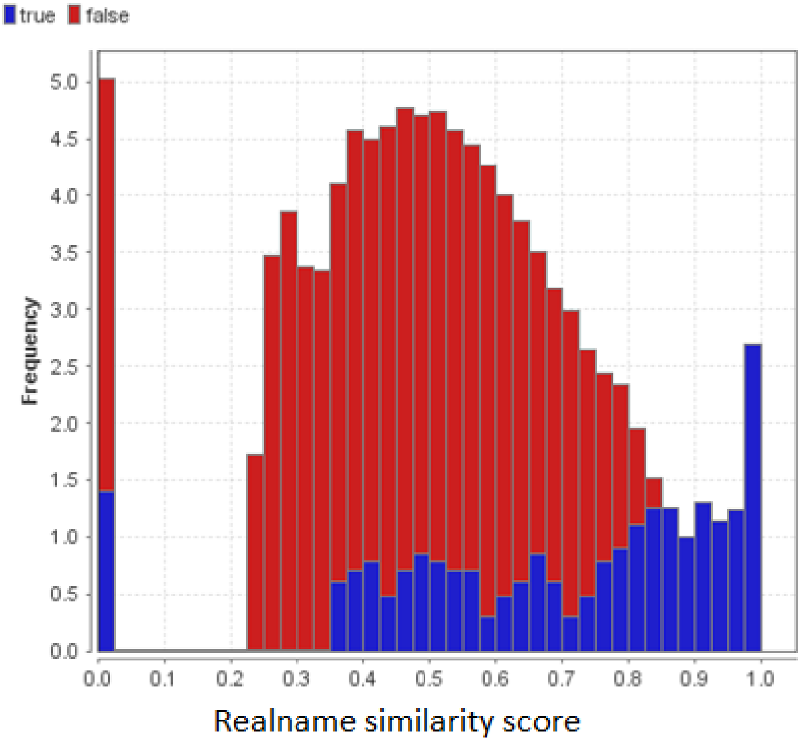

Figure 4. Histogram of RN scores

Figure 5 shows that distribution for correct matches is not highly stacked at higher scores. Only quarter of the positive matches have similarity score close to 1. Since users have to choose a unique username on each social network, some may choose slightly different but very similar to the username used on the other social network because that username is already taken by another user. Threshold value is 0.85 and TP rate is 49%.

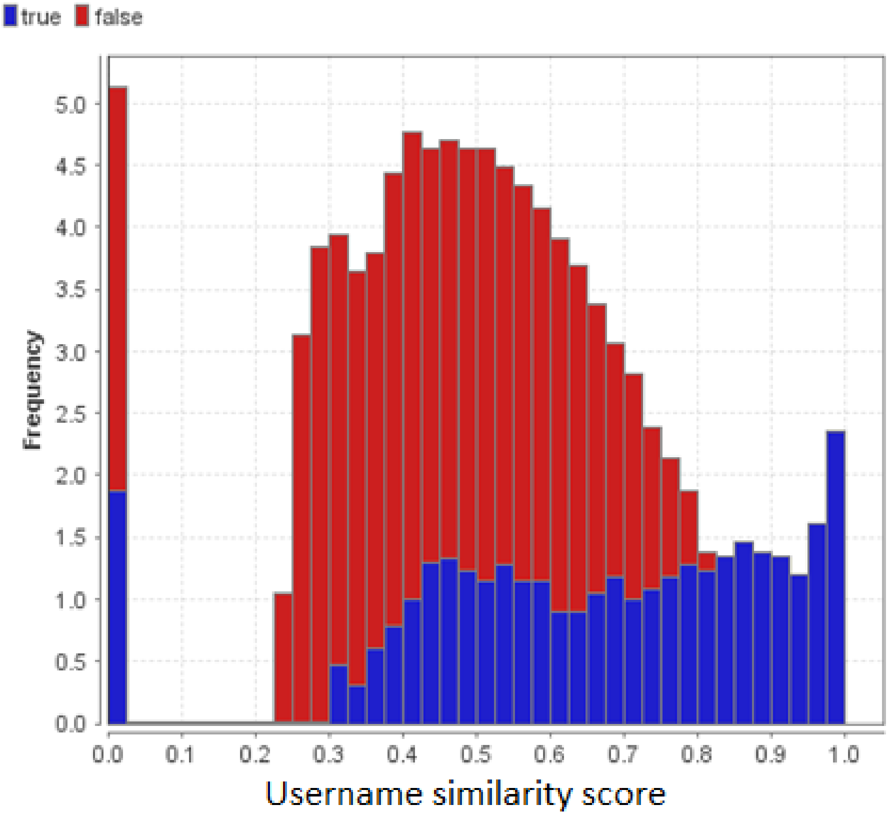

Figure 5. Histogram of UN scores

Figure 6 and Figure 7 shows that distributions for correct matches are not skewed and shows that people choose slightly different usernames on a social network than real names used on the other network. For CN1 threshold value is 0.78 and TP rate is 30%. For CN2 threshold value is 0.82 and TP rate is 36%.

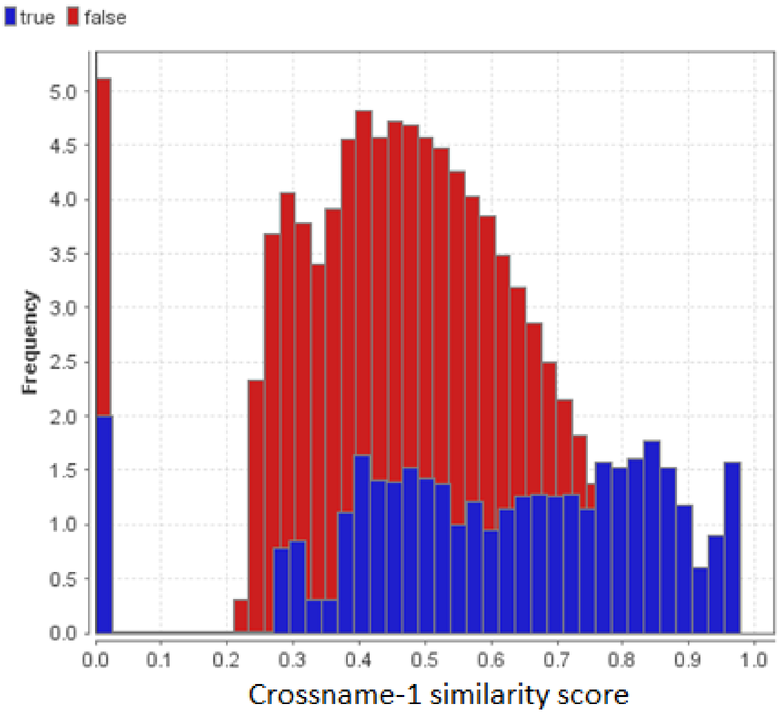

Figure 6. Histogram of CN1 scores (from Twitter UN to Facebook RN)

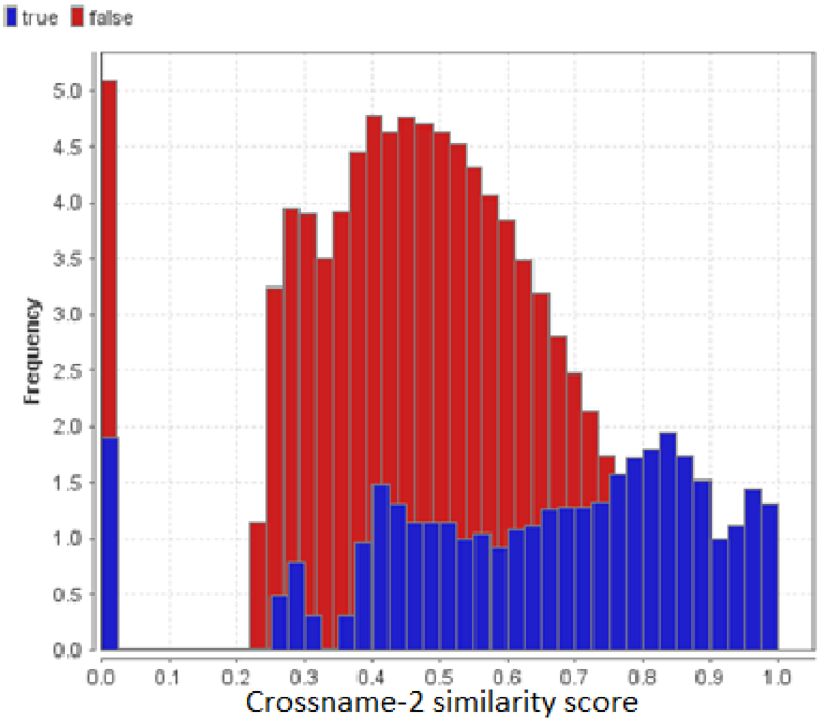

Figure 7. Histogram of CN2 scores (from Twitter RN to Facebook UN)

Figure 8 shows that text similarity between two profiles of the same person on two SN is very low and it is hard to match profiles using this feature except a few person which shared the exact same status updates in both accounts. Threshold value is 0.28 and TP rate is 5%

.

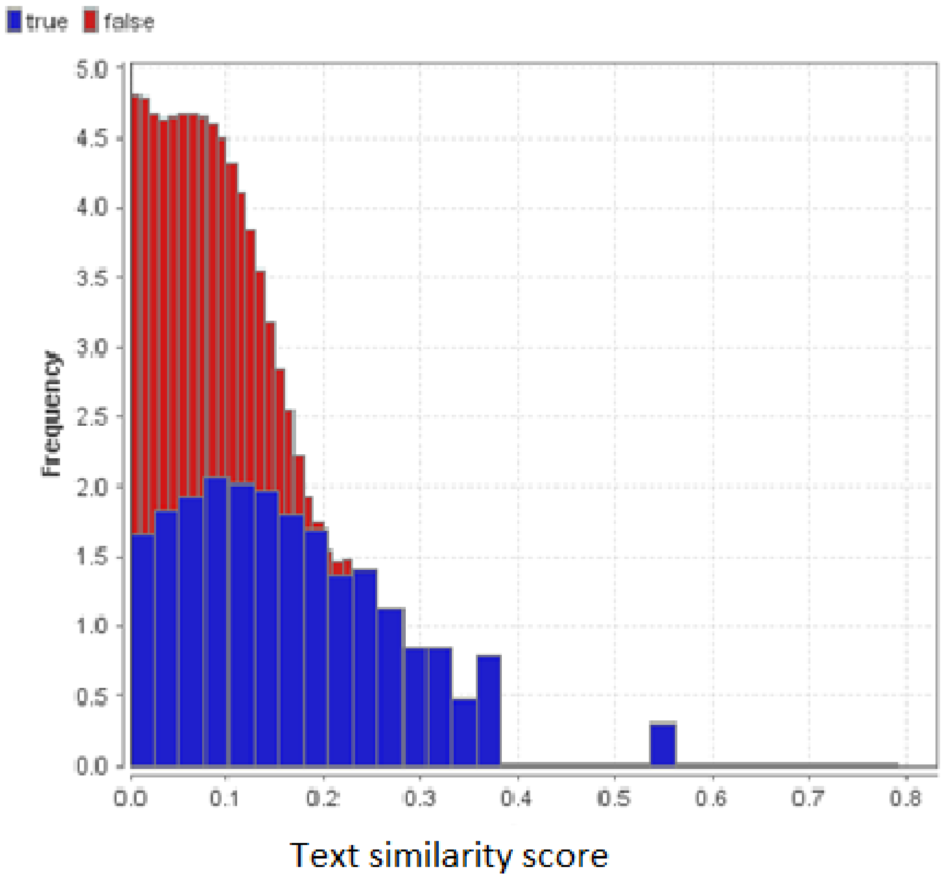

Figure 8. Histogram of TEXT scores between Facebook and Twitter posts

Figure 9 shows ROC curves for each feature. X axis is for FP rate in log scale and Y axis is for TP rate. RN has highest TP rates and Text has the lowest TP rates as expected. That means if only one feature is available at a time classifier finds 71% of the profiles correctly by their real names and only 5% can be found by comparing their posts for a FP rate of 10^{-5}. If all features are combined then more profiles can be matched.

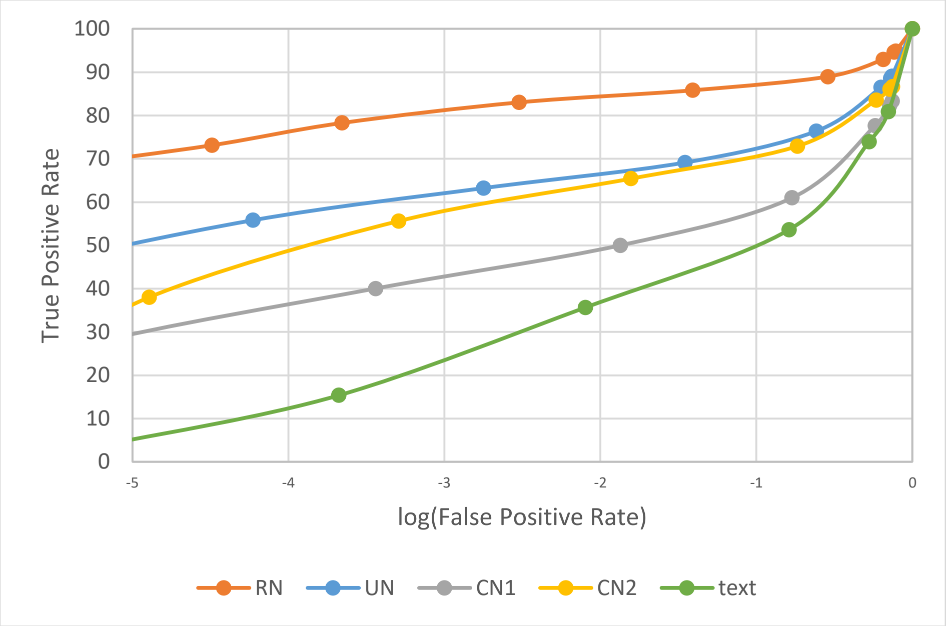

Figure 9. ROC curves for each feature

6.1.Combining Features

After seeing individual performances of each feature, all features are combined to match user accounts with the help of different data mining algorithms. Algorithms are implemented on sampled dataset because it is not possible to run SVM and Backpropagation on original dataset (which has 4273 TPs and more than 18 million FPs) with limited computational resources in the study laboratory. There are 4273 TPs and 10000 FPs in sampled dataset.

Table 10 shows performance scores of different classification algorithms for 0.95 precision. First group which is under the cell RN, shows performance scores of the algorithms using only RN feature. On original dataset Logistic Regression, Naïve Bayes and Decision Tree algorithms shows 0.70 recall value for RN and in sampled dataset it increases to 0.85. This is due to sampling FPs which changes the class boundaries and gets higher recall value with fixed precision.

Table 10. Performance scores of classification algorithms on sampled dataset

	RN		RN+UN+CN1+CN2	
Algorithm	**Recall**	**AUC**	**Recall**	**AUC**
Logistic Regression	0.85	0.914	0.85	0.935
Naïve Bayes	0.85	0.914	0.89	0.955
KNN 1	0.84	0.935	0.90	0.946
KNN 5	0.84	0.94	0.91	0.963
Decision Tree	0.84	0.918	0.91	0.957
Backpropagation	0.85	0.925	0.92	0.963
SVM	0.83[1]	0.913	0.91[2]	0.954

There is no significant difference between algorithms when single feature is used for classification. However when all name features are added, Decision Tree, Backpropagation and SVM have the highest recall values among all as expected. Classification boundaries and expected performances of each algorithm is inspected in Appendix A.

Table 11 shows effect of sampling and improvements by combining all name based features with decision tree algorithm. On the full dataset recall value is 0.82 where recall value is 0.92 on sampled dataset.

Table 11. Comparison of recall values due to sampling

Features	**Full Dataset**	**Sampled dataset**
RN	0.70	0.85
RN+UN+CN1+CN2	0.82	0.91

Since TEXT feature is only available for subset of the profiles, improvement of adding TEXT feature is seen on this smaller dataset (849 users). Recall is increased from 0.70 to 0.79 because subset is smaller than full dataset (4273 users). After adding TEXT as a new feature, as seen from Table 12, the recall is increased from 0.87 to 0.88 which shows poor improvement as a feature.

[1] Precision 0.993 SVM.

[2] Precision 0.995 SVM.

Table 12. Recall values of TEXT subset

Features	TEXT Subset
RN	0.79
RN+UN+CN1+CN2	0.87
RN+UN+CN1+CN2+TEXT	0.88

6.2.Improving Performance of TEXT Feature

TEXT feature has the worst performance alone and adding it as a new feature does not improve the matching schema significantly. In this section, factors that affect the TEXT similarity score and improvement methods are evaluated.

TEXT similarity depends on the terms (words) users post on Twitter and Facebook. If number of terms in common at both accounts increase then similarity score increases between two profile. If a term is repeated many times at a profile TF value increases for that term. However if a term is repeated at many profiles and if it is common for many profiles then IDF value decreases for that term. Therefore high similarity score for two profiles of the same person depends on two things:

- User should post similar things in both accounts. Interest area of the user in both networks and comments left by the user should belong to similar topics. If number of common terms is very low then similarity between profiles will be very low.
- User should post unique things in both accounts. If user posts similar things like other users in the same network then TF-IDF vector will be similar to other users' which makes its profile harder to distinguish from other users'.

Figure 10 and 11 shows the relationship between the number of terms posted by users and TEXT scores among two profiles of the same users in both networks. X axis is for term counts and Y axis is for TEXT similarity score for correct matches

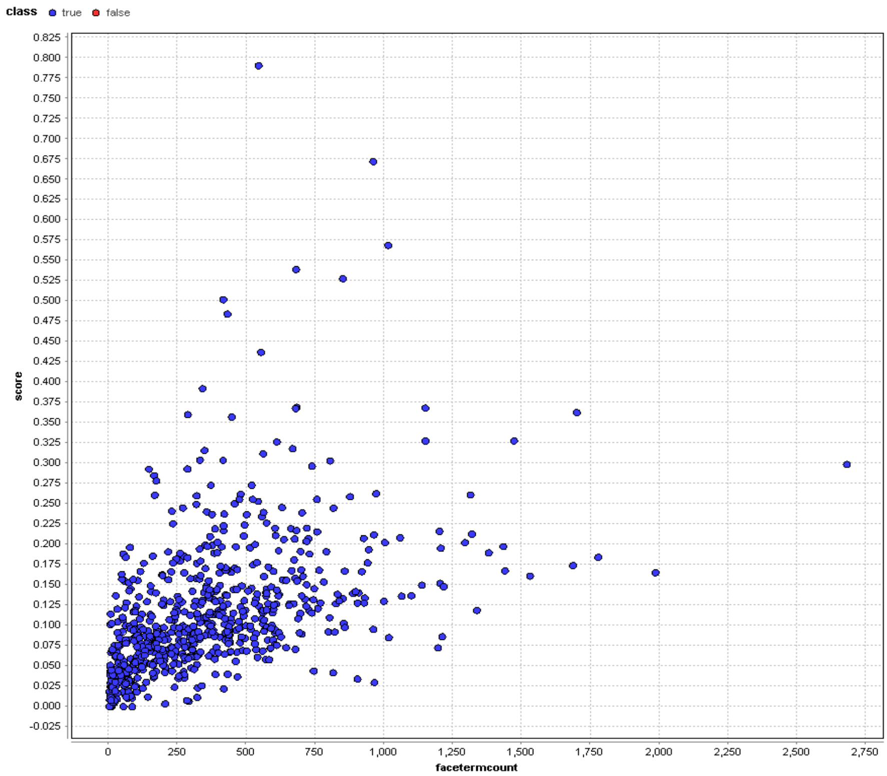

Figure 10. Text similarity scores (y-axis) and Facebook Terms (x-axis) for correct matches

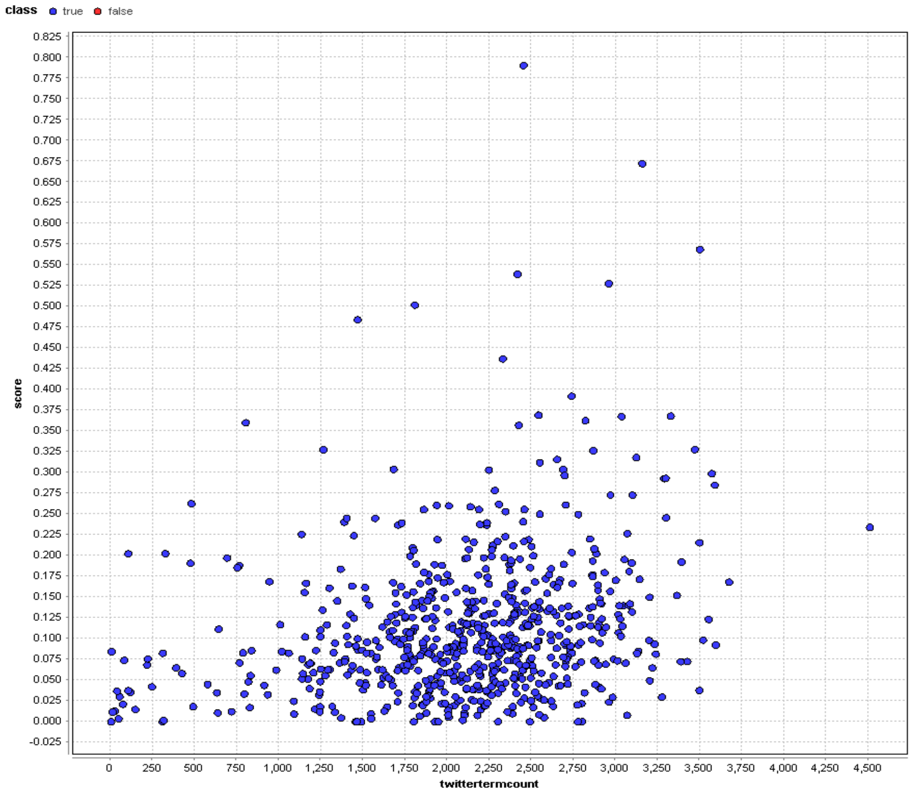

Figure 11. Text similarity scores (y-axis) and Twitter Terms (x-axis) for correct matches

Table 13 shows the linear correlation matrix between TEXT similarity score, Facebook term count and Twitter term count. It is seen that correlation between similarity score and Facebook term count is higher than Twitter term counts. That means TEXT similarity scores are more dependent to Facebook term counts rather than Twitter term counts. So the more terms a user posts on Facebook, higher TEXT similarity scores it will get. Also there is less linear correlation between term counts of the same people on two networks. That means users post different number of terms in different networks.

Table 13. Linear correlation matrix for correct matches

Attributes	Score	Twitter Term Count	Facebook Term Count
Score	1	0.255	0.50
Twitter Term Count	0.25	1	0.22
Facebook Term Count	0.50	0.22	1

Table 14 shows correlation between term counts and TEXT similarity score of matches between different people. Low correlation values in this table tells us, increasing in term counts does not lead to higher score for false matches. As users post more in a network, their profile will not be similar to other profiles of different users in the same network.

Table 14. Linear correlation matrix for incorrect matches

Attributes	Score	Twitter Term Count	Facebook Term Count
Score	1	0.252	0.179
Twitter Term Count	0.252	1	0.262
Facebook Term Count	0.179	0.262	1

Correlation results tell that probability of a user will be detected by TEXT is slightly dependent to number of Facebook terms that is posted by the user. Number of terms posted at Twitter profile is not significant since increase in terms increase TEXT score of false matches as same amount. One of the reasons can be the different natures of the networks and users choose using two networks for different purposes. In Twitter they post more tweets and talk about recent and popular topics. Twitter encourages users to tweet about popular topics by showing a list-pane in the home page called "Trending topics”. It is updated in real time and shows global and local topics and hashtags which are derived from most popular and most talked about news, events, people or places. As more users post about same topics in Twitter their TEXT documents will be similar and distinguishing the user from others will be harder. Unlikely the users of Facebook post less and talk about less popular topics than Twitter topics. This make their TEXT documents more discriminative and increase in term count leads to higher TEXT scores for true matches.

Figure 12 and 13 show histogram of term counts of the profiles. X axis denotes term counts and Y axis denotes the number of profiles. As seen from Figure 12, distribution of Facebook Term counts in the dataset is skewed right and big part of the users has very low term counts. Maximum term count is 2680 and average is 354. This situation is due to people on Facebook post less status updates which are publicly accessible. Privacy settings of Facebook let users choose to post publicly or more restrictively for each status update. If user does not change the privacy preference and keeps to share status update only with his friends then it will not be accessible by the crawler and term count will be smaller. However, in dataset Twitter profiles are public since in preprocessing step, private accounts are excluded from analysis. As seen from Figure 13, distribution is symmetric and average term count is 2107 for Twitter profiles.

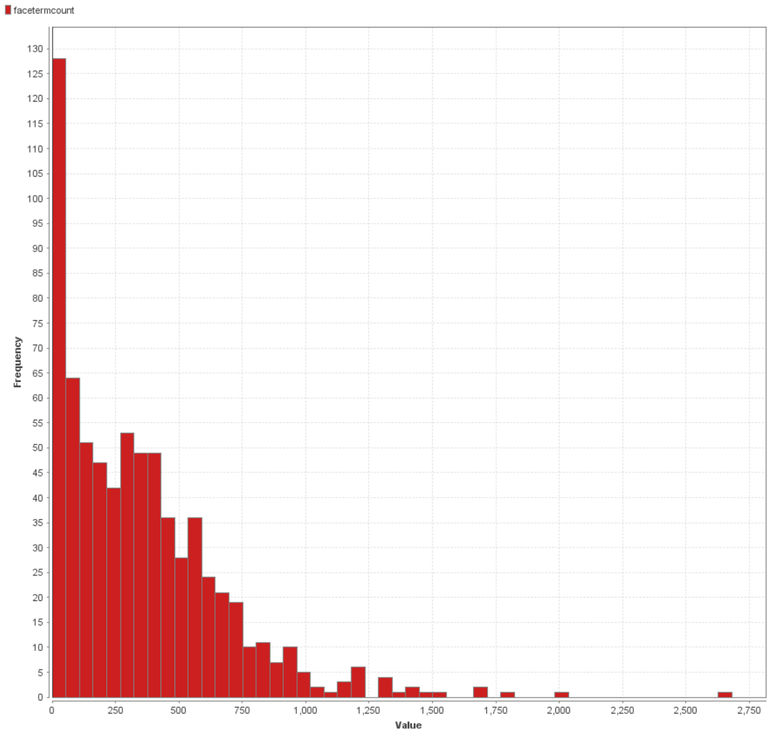

Figure 12. Histogram for Facebook Term counts

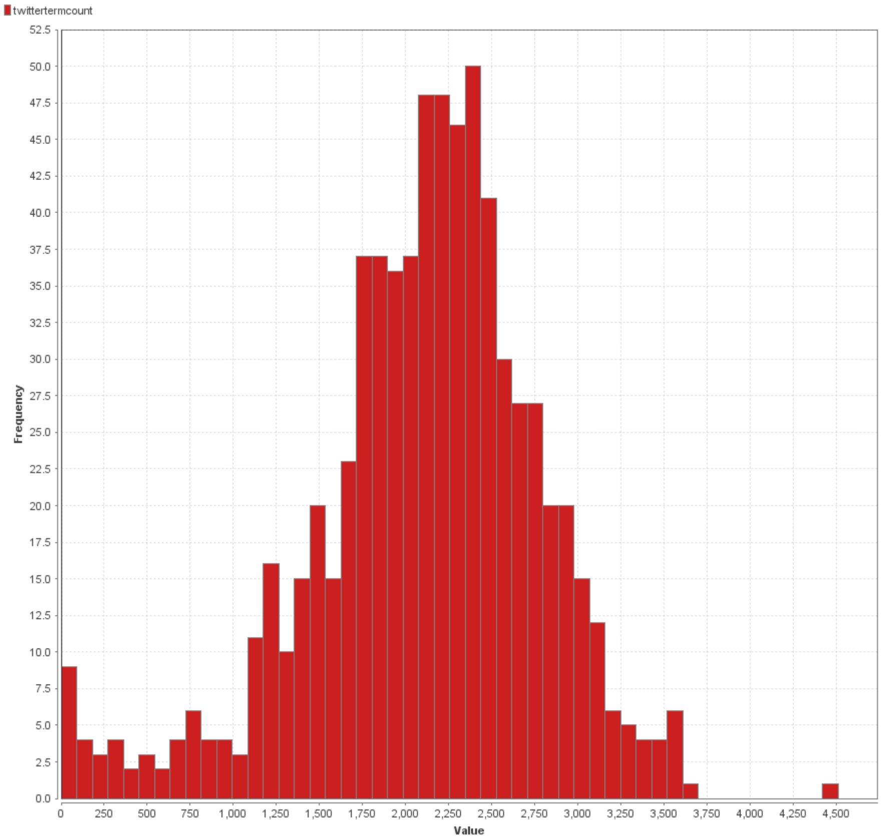

Figure 13. Histogram for Twitter Term counts

Assume as term count increases for a profile in SN, its TEXT similarity score with other profile on the other SN increases and matching possibility of two accounts increases. Therefore limiting matching scheme by term counts may increase the performance. Discarding profiles with low term values and processing only profiles with the minimum amount of term counts, reduces the number of candidate accounts for each profile and eliminates the true matches that can’t be detected with high precision values.

First approach is discarding subset of profiles with low term counts in destination SN which is Facebook in this study. Matching scheme will not include subset of Facebook profiles for comparison, therefore profiles in source SN will be compared with a subset of profiles at destination SN. Second approach is discarding subset of profiles with low term counts in source SN which is Twitter.

Figure 14 shows recall values for TEXT feature after filtering Facebook profiles with low term counts. X axis shows the minimum term counts for the profiles in destination SN. Initially recall value is 2% when all Twitter profiles are compared with all Facebook profiles. As Facebook profiles with low term counts are excluded from comparison, recall value increases. If all Twitter profiles are compared with Facebook profiles with minimum term count of 1000 then recall value is 20% where 29 Facebook profiles are in the destination dataset and 6 of them are classified correctly.

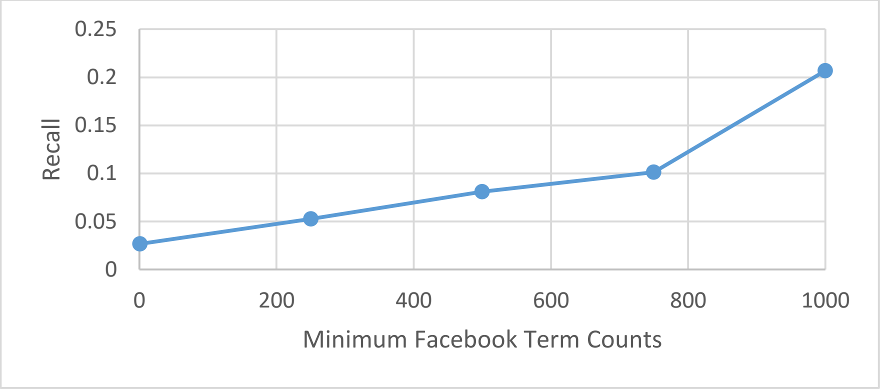

Figure 14. Recall values after filtering Facebook profiles with low term counts

Figure 15 shows recall values for TEXT feature after filtering Twitter profiles with low term counts. X axis shows the minimum term counts for the profiles in source SN. When Twitter profiles with low term counts are excluded from comparison, recall value increases as well. If all Facebook profiles are compared with Twitter profiles with minimum term count of 3000, then recall value is 16% where 55 Twitter profiles are in the destination dataset and 9 of them are classified correctly.

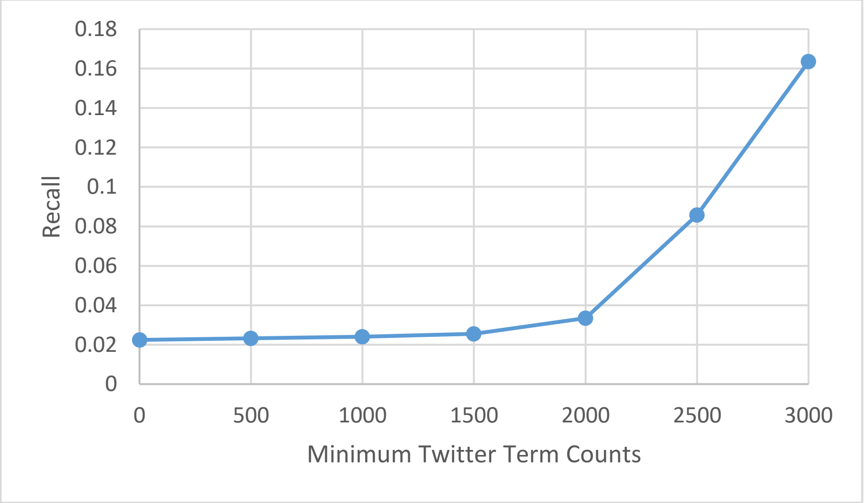

Figure 15. Recall values after filtering Twitter profiles with low term counts

6.3.Top-k Approach

In previous sections the problem is a classification problem where each instance to be classified is a comparison between any profile pair in source SN and destination SN. However original problem is a matching problem where each profile in source SN has one corresponding profile in destination SN. In the classification problem each instance is classified independently from each other. Even a correct match is classified as true match, many false matches of the same profile can also be classified as true match which increases the FP rate and reduces the recall value. If most similar profiles are classified as true match only, many false positives can be eliminated and matching rate can increase. Therefore new methodology is as follows:

- Comparing each profile in source SN with each profile in destination SN.
- For each profile in source SN, sort similarity scores of matching with other profiles in destination SN.
- Choose top-k similar profiles in sorted list.
- If correct match is in the list, then two profiles of the same individual is detected successfully. Else, all profiles in the list are false matches and correct profile is not detected.

In section 6.1, the sampled data set has 4273 TPs and 10000 FPs which means there are less than 2.5 false matches for each profile on the source network. However it is not possible to apply top-k approach because top-3 achieves 100% matching rate already. In this section there are 4273 TPs and 80000 FPs in sampled dataset which means there are more than 18 candidate profiles in destination network for each profile in the source network.

Figure 16 shows the match rate for different k values. Four name features (RN, UN, CN1, and CN2) are combined to get probability score with Logistic Regression. In the classification problem Logistic Regression has 0.81 recall on sampled dataset. If most similar profiles (top-1) are selected as true matches then 91% of the profiles can be detected. If k value increases matching rate increases as well. If k value is 10 then 96% of the profiles can be matched within 10 most similar profiles.

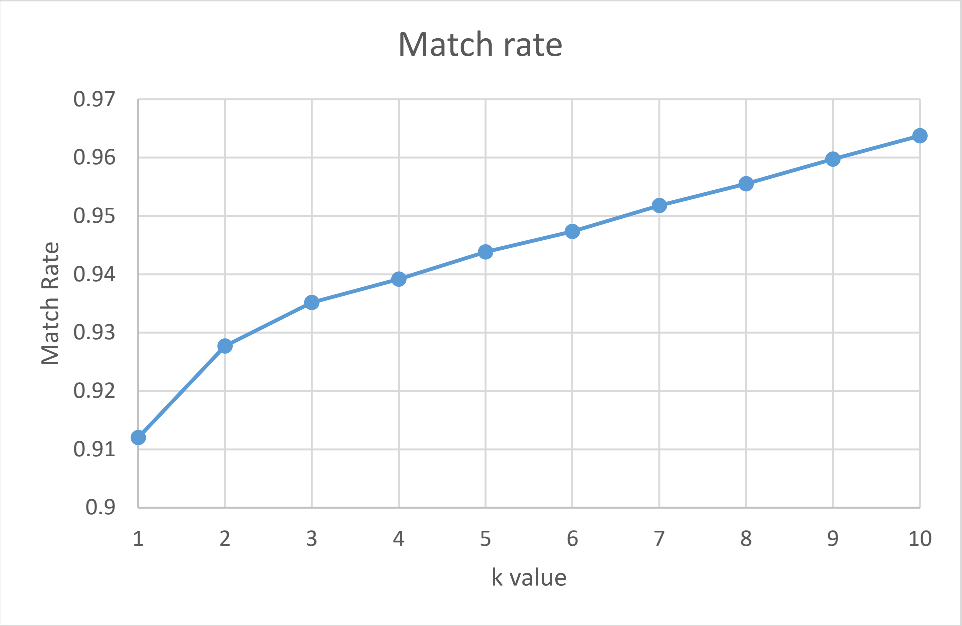

Figure 16. Match rate of profiles after choosing top-k similar profiles

CHAPTER 7

CONCLUSION

More than 97.000 Google+ profiles are crawled to gather ground truth data which consist of URLs of different profiles on various SNs that are shared by the users willingly. Most provided URLs belong to YouTube, Twitter and Facebook respectively. Facebook and Twitter profiles of the same users are chosen for this study because name based attributes and status updates are available in both networks. Among 97.000 users only 4273 of them have valid profiles in both networks and status updates of only 848 users is accessible due to privacy settings.

Three attributes are chosen for analysis which are publicly available and in common for Facebook and Twitter. They are real name, user name and status updates and five features are derived from these attributes which are RN, UN, CN1, CN2 and TEXT.

Results show that;

- Users use very similar real names and user names in their Facebook and Twitter profiles which makes name based attributes the most consistent ones. These attributes are also the most discriminative attributes because users have different real names and choose different user names and it makes them distinguishable from each other. However in larger scale or in a real life application attributes will become less discriminative because number of profiles to be compared will increase and there will be many users share the same or similar attribute values in the dataset. If there are other attributes in common among SNs that are consistent and publicly available, then the possibility of matching increases.
- If each feature is used alone for comparison and matching profiles, RN is the best feature to match profiles with 70% recall and 95% precision. UN has 48%, CN1 has 22%, CN2 has 21% and TEXT has 2% recall value.
- If features are combined together to get probability scores of matching, then recall value increases from 70% to 82%. In the original matching problem there is always one matching profile in the destination network. Knowing that classifying many false matches of the same profile decreases precision. With top-k approach probability scores are used to get most similar profiles in destination network for each profile in source network, then matching rate increased Matching only most

similar profiles (top-1) makes matching rate 91% and 96% of the profiles can be matched if most 10 similar profiles (top-10) are gathered.

- (**Contribution 1&2**) TEXT feature is the least consistent feature and its relation with how much a user posts on two networks is analyzed to improve its performance. Analyze on total terms posted by the users gives some insights about how different two networks are used. It is seen that users share more status updates in Twitter than Facebook. Also total terms posted by the same user on different networks are not correlated. TEXT similarity score between the profiles of the same person is more dependent to total terms posted by the user in Facebook than total terms in Twitter. Excluding profiles with low term counts increased recall value from 2% to 20%.
- (**Contribution 3**) Different classification algorithms are also analyzed to determine the most suitable model for evaluation. When each feature is used alone, different algorithms showed similar performances however after all features are combined to classify instances, each algorithm produced different decision boundaries. SVM, Backpropagation and Decision Tree algorithms are the most suitable ones to perform classification problem on matching profiles.
- As compared to related studies, performance of matching accounts is similar and name based attributes have the best performance as expected.

Results show that it is possible to match two profiles that belong to same user using publicly accessible information. An attacker can detect high similarity between two profiles and reveal identity of the user which poses threat to the privacy of the people on the Internet. Matching possibility of two profiles of the same user on different networks depends on availability, consistency and discriminability of the common attributes across networks and number of candidate profiles that are used in comparison.

7.1.Recommendations to the Users

Users on internet must read privacy policies of the service providers and should know which data about the user is collected, used and shared with third parties. Users must be aware of that free services are usually offered to learn user behaviors and for marketing purposes.

Each SN has different nature and content sharing mechanism. Privacy settings and configurations can help the users to protect their privacy however some attributes of the profiles are can be public and can be accessible by search engines, crawlers and other members of the networks. If user wants to be undetected matching schemas such as mentioned in this study, firstly public attributes of the profiles must be determined than they should be differentiated.

The most effective way to avoid detection is choosing different real names and usernames since they are the most discriminative and consistent attributes. Also choosing very common names makes the profile very similar to unrelated profiles and harder to match. Usually visibility of other attributes such as personal information, age, sex and friendship lists can be managed with privacy settings and can be unreachable by public. However profile photos are usually publicly accessible as default. Uploading different photographs or not uploading at all can avoid such detection. However user must be aware of that many service providers at different layers (internet service providers, application owners, content providers and other third parties) have access some other user identifiers such as IP address, activity time patterns, interest areas, time zone and other information about activities. This information can be used to construct more complete online identity of the user if this linkable data is associated when companies merge, data is shared or exploited.

CHAPTER 8

FUTURE WORK

Performance of a matching scheme depends on the availability, consistency and discriminability of the attributes that are chosen among SNs. Each network has different attribute and nature, therefore choosing different SNs will require analysis of different attributes. Some of the attributes can be used for matching are photographs, friendship graph, locations, genders, URLs, time zones and interest areas.

For Facebook and Twitter analysis, photograph similarity can be added to the matching scheme since profile pictures for both SNs are available.

In TEXT similarity section, publish time of the status updates are not included in the analysis. However comparing status updates that are published in the same time intervals such as one hour, one day, one week etc. may be more significant. Giving attention to publish times may increase the performance of TEXT similarity.

In Twitter, users are limited with 140 characters per tweet but in Facebook there is no such limitation. It is found that TEXT similarity is more dependent to terms count of Facebook status updates. The reason of this dependency can be character limit on Twitter.

REFERENCES

[1] Danah M. Boyd and Nicole B. Ellison, "Social Network Sites: Definition, History and Scholarship," Journal of Computer-Mediated Communication, 2007.

[2] Xi Chen and Shuo Shi, "A Literature Review of Privacy Research on Social Network Sites," in Proceedings of the 2009 International Conference on Multimedia Information Networking and Security - Volume 01 (MINES '09), Vol. 1. IEEE Computer Society, Washington, DC, USA, 2009.

[3] S. Goodson, "If You're Not Paying For It, You Become The Product," Forbes, 5 March 2012. [Online]. Available: http://www.forbes.com/sites/marketshare/2012/03/05/if-youre-not-paying-for-it-you-become-the-product/#43c79c19b445.

[4] Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira, "Your browsing behavior for a big mac: economics of personal information online," in Proceedings of the 22nd international conference on World Wide Web (WWW '13) ACM, New York, 2013.

[5] Erika McCallister, Tim Grance, and Karen Scanfone, "Guide to protecting the confidentiality of personally identifiable information (PII)," NIST, April 2010. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf. [Accessed 2016].

[6] Balachander Krishnamurthy and Craig E. Wills., "Characterizing privacy in online social networks," in Proceedings of the first workshop on Online social networks, New York, NY, USA, 2008.

[7] Balachander Krishnamurthy and Craig E. Wills, "On the leakage of personally identifiable information via online social networks," in roceedings of the 2nd ACM workshop on Online social networks, New York, NY, USA, 2009.

[8] Delfina Malandrino, Vittorio Scarano, "Privacy leakage on the Web: Diffusion and countermeasures," Computer Networks, vol. 57, no. 14, pp. 2833-2855, 2013.

[9] L. Sweeney, "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh, 2000.

[10] L. Sweeney, "k-anonymity: a model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557 - 570, 2002.

[11] H. Wang, Innovative Techniques and Applications of Entity Resolution, Hershey, PA, USA: IGI Global, 2014.

[12] P. Christen, Data Matching, Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection, Springer-Verlag Berlin Heidelberg, 2012.

[13] Andreas Pfitzmann and Katrin Borcea-Pfitzmann, Lifelong Privacy: Privacy and Identity Management for Life, Berlin: Springer-Verlag, 2009.

[14] Almishari, Mishari and Gene Tsudik, "Exploring linkability of user reviews," in ESORICS 2012, Pisa, Italy, 2012.

[15] O. Peled, M. Fire, L. Rokach and Y. Elovici, "Entity Matching in Online Social Networks," in Social Computing (SocialCom), Alexandria, VA, US, 2013.

[16] R. Soltani and A. Abhari, "Identity matching in social media platforms," in Performance Evaluation of Computer and Telecommunication Systems (SPECTS), Toronto, 2013.

[17] Ye Na, Zhao Yinliang, Dong Lili, Bian Genqing, Enjie Liu and G. J. Clapworthy, "User identification based on multiple attribute decision making in social networks," China Communications, vol. 10, no. 12, pp. 37-49, 2013.

[18] Goga, O., Perito, D., Lei, H., Teixeira, R., & Sommer, R, "Large-scale correlation of accounts across social networks," University of California at Berkeley, Berkeley, California, 2013.

[19] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P. Gummadi, "On the Reliability of Profile Matching Across Large Online Social Networks," Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1799-1808, 2015.

[20] "Facebook Developer Graph API Rerefence," [Online]. Available: https://developers.facebook.com/docs/graph-api/reference/user. [Accessed 20 March 2014].

[21] "User Object," Twitter, [Online]. Available: https://dev.twitter.com/overview/api/users.

[22] Twitter, "Tweets," [Online]. Available: https://dev.twitter.com/rest/reference/get/search/tweets. [Accessed 2016].

[23] Twitter, "Rate Limiting," [Online]. Available: https://dev.twitter.com/rest/public/rate-limiting. [Accessed 2016].

[24] Perito, Daniele, Claude Castelluccia, Mohamed Ali Kâafar and Pere Manils, "How Unique and Traceable Are Usernames?," in Privacy Enhancing Technologies, Waterloo, ON, Canada, 2011.

[25] Cohen, William W., Pradeep D. Ravikumar, and Stephen E. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks," in IIWeb, Acapulco, Mexico, 2003.

[26] S. Wilson, "HTMLCleaner," [Online]. Available: http://htmlcleaner.sourceforge.net/. [Accessed 2016].

[27] A. Green, "140dev Streaming API Framework," 140dev, [Online]. Available: http://140dev.com/free-twitter-api-source-code-library/. [Accessed 2016].

[28] J. Mallison, "Simple PHP Wrapper for Twitter API v1.1 calls," [Online]. Available: https://github.com/J7mbo/twitter-api-php. [Accessed 2016].

[29] M. Bowler, "HtmlUnit," Gargoyle Software, [Online]. Available: http://htmlunit.sourceforge.net/. [Accessed 2016].

APPENDIX A

DECISION BOUNDARIES

In this section different data mining algorithms are compared by visualizing their classification boundaries to see the most suitable algorithms for the matching problem. Each algorithm shows similar performance when using single feature at a time; however when features are used together for classification, the performance between them differs due to their classification boundaries. The shape of the boundaries depends on the algorithm, features and similarity metrics. And the shape of the boundaries determine the TP and FP rates.

Since there are five features, true matches and false matches are distributed in a 5-dimensional space. Ideal classifier finds boundaries that separates true matches and false matches such as TP rate is 100% and FP rate is 0%. Each classification algorithm finds different boundaries for separation and that effects the recall and precision values. Distribution of each feature and selected algorithm determines the boundaries and performance of the classifier. Therefore choosing the most suitable algorithm that draws the best decision boundary is important. In this section different algorithms are inspected to understand how they performed for classification of matches. Visualization of distributions and boundaries are shown on 2 dimensional space to give insights about higher dimensions.

In Figure 17 and 18, distribution of similarity scores are represented. Green colors represent false matches and red colors represent true matches. Scores are between 0 and 1 where 1 is the highest similarity and 0 is the lowest similarity. In Figure 10 most discriminative features are shown, where x axis is for RN and y axis is for UN. In Figure 11 most discriminative feature RN and least discriminative feature are shown, where x axis for TEXT and y axis is for RN.

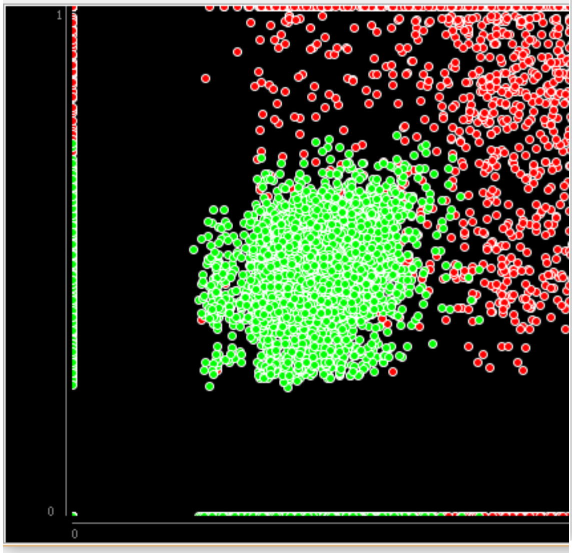

Figure 17. Distribution of RN (x axis) and UN (y axis)

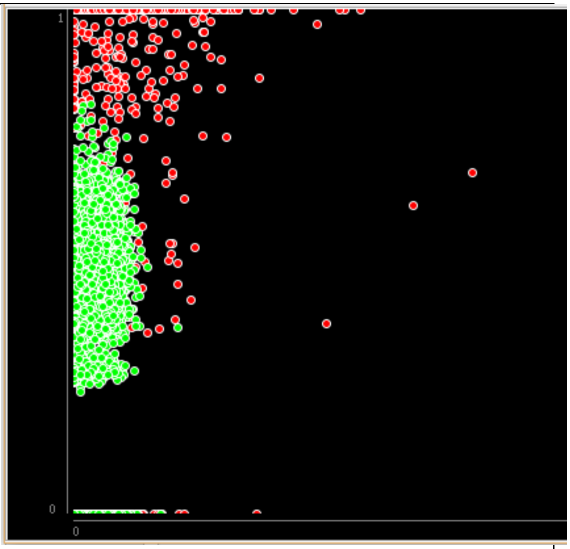

Figure 18. Distribution of RN (y axis) and TEXT (x axis)

Figure 19 to 30 shows results of different classification algorithms for the distribution in Figure 17.

In Figure 19, Logistic Regression finds a linear classification boundary however for users whose real names are different but use similar user names (red colors on the upper left in in Figure 17) will not be detected and they will be classified as false match. Also most of the users whose real name is similar but use different user names (red colors on the lower right in Figure 17) will not be detected. Only users that use similar real names and user names will be detected by the Logistic Regression.

In Figure 20, Naïve Bayes find a similar but smoother curve than Logistic Regression. It will increase the recall value however it is still has a bad boundary since many true matches with high scores will not be classified as positive.

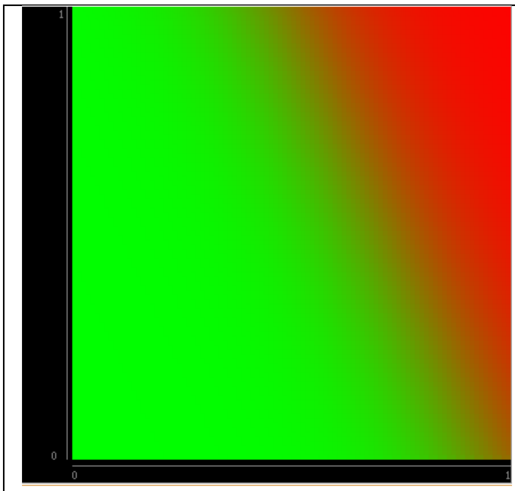

Figure 19. Boundary of Logistic Regression for RN and UN

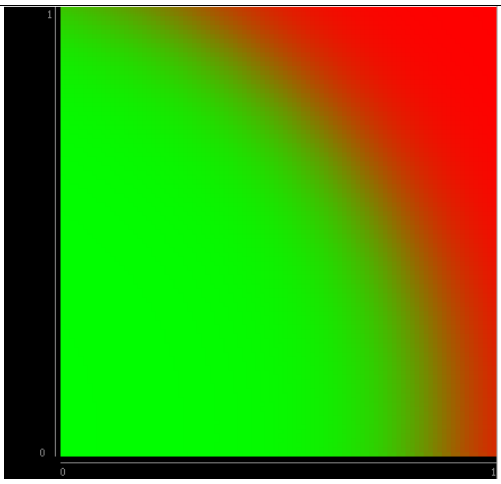

Figure 20. Boundary of Naïve Bayes for RN and UN

In Figure 21 and 22, boundaries for nearest neighbor algorithm are shown for k values 1 and 5 respectively. KNN determines if a match is positive or negative by looking up nearest matches in the training data. It is seen that there are some red areas inside the green area which does not make sense. However it finds better boundaries since true matches with high real name or user name similarity scores will be classified as positive by KNN.

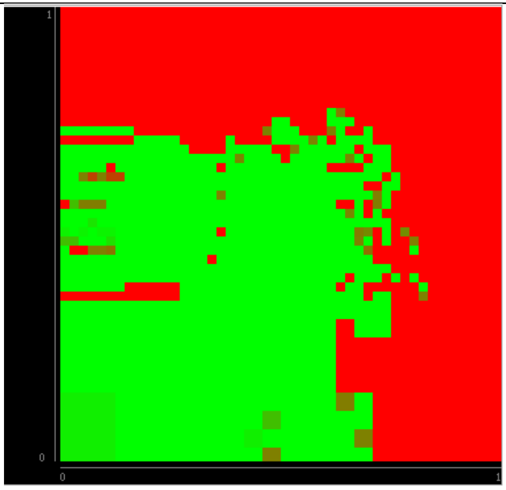

Figure 21. Boundary of KNN-1 for RN and UN

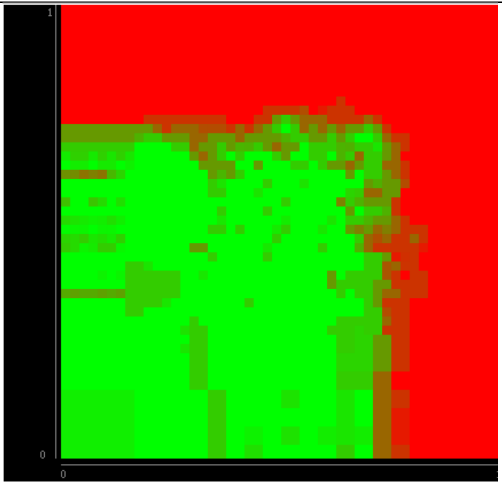

Figure 22. Boundary of KNN-5 for RN and UN

In Figure 23, Decision Tree finds the sharpest boundaries between two classes. There are two main branches in the tree. First one is for RN. Matches with RN score higher than 0.87 are classified as true. Second branch is for UN. Matches with UN score higher than 0.89 are classified as true.

In Figure 24, Backpropagation shows the most suitable decision boundary for the distribution in Figure 17. It is a smooth boundary and does not suffer from overfitting unlike Decision Tree.

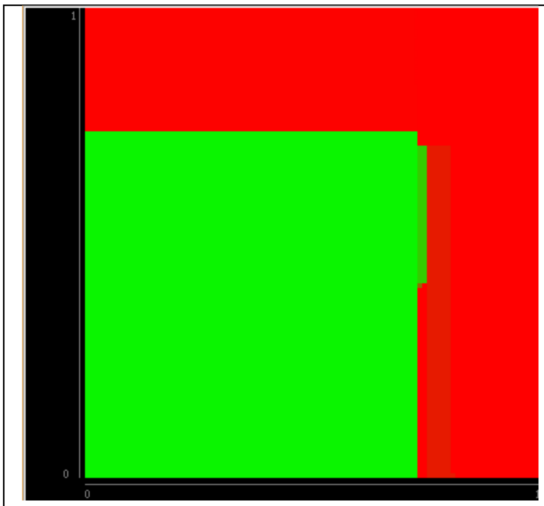

Figure 23. Boundary of Decision Tree for RN and UN

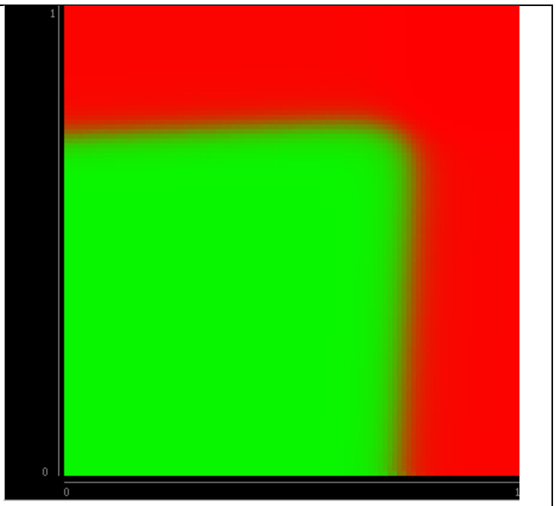

Figure 24. Boundary of Backpropagation for RN and UN

Figure 25 to 30 shows results of different classification algorithms for the distribution in Figure 18 where TEXT scores are very low between matching accounts. Recall value is very low for TEXT feature and there are a few true matches with low RN scores and high TEXT scores. It is expected that TEXT feature will improve performance of the matching scheme by classifying true matches on lower right corner.

In Figure 25 part of the true matches with very high RN scores are above and some part are below the decision boundary. Also some of the true matches with low RN and moderate TEXT scores which are expected to improve performance of the matching scheme, are below the boundary.

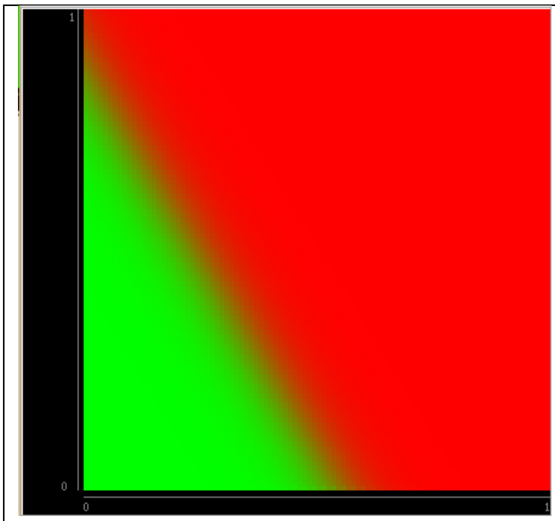

Figure 25. Boundary of Logistic Regression for RN and TEXT

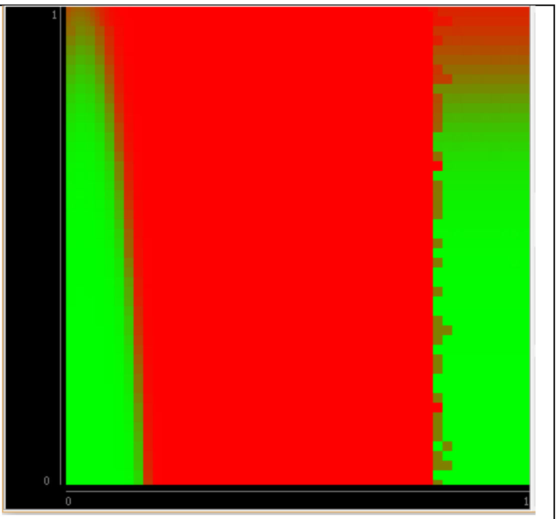

Figure 26. Boundary of Naïve Bayes for RN and TEXT

In Figure 26, Naïve Bayes classifier draws a noisy boundary since training set suffers from lack of true matches with high TEXT score. Figure 27 and 28 also shows noisy boundaries because of false matches with high TEXT score. Since there is a data imbalance between true and false matches some areas can be decided to negative class even there are some positive matches.

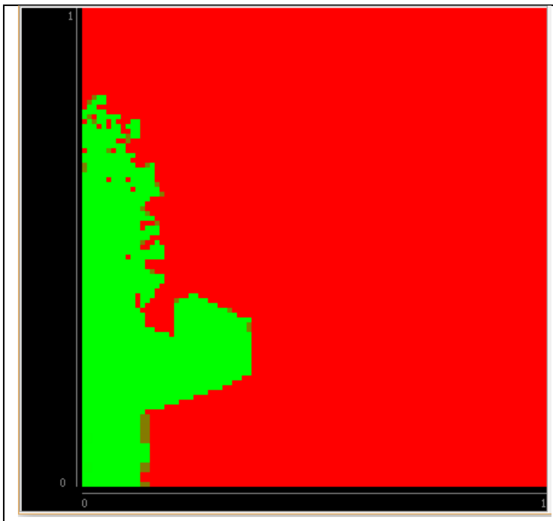

Figure 27. Boundary of KNN-1 for RN and TEXT

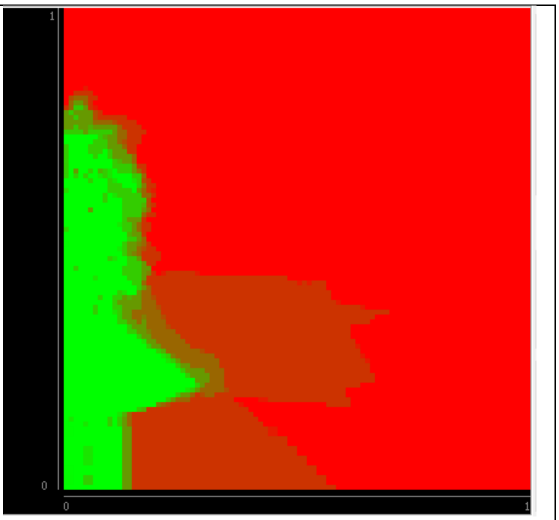

Figure 28. Boundary of KNN-5 for RN and TEXT

Figure 29 and 30 shows more precise boundaries than others however Decision Tree branches more than two times which leads to an overfitting model. Decision boundary for Backpropagation is the best among others because it would give better results when RN and TEXT scores are high for true matches.

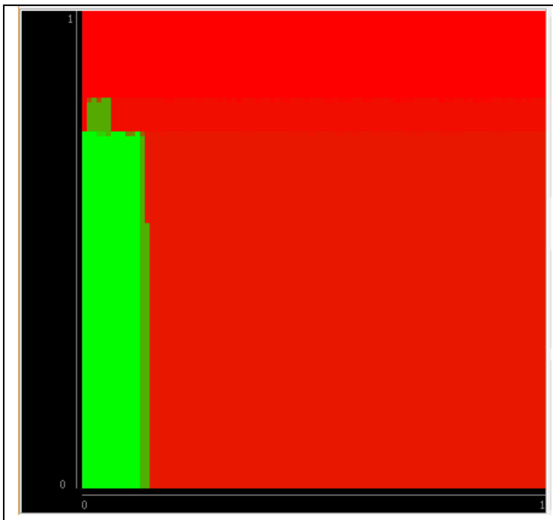

Figure 29. Boundary of Decision Tree for RN and TEXT

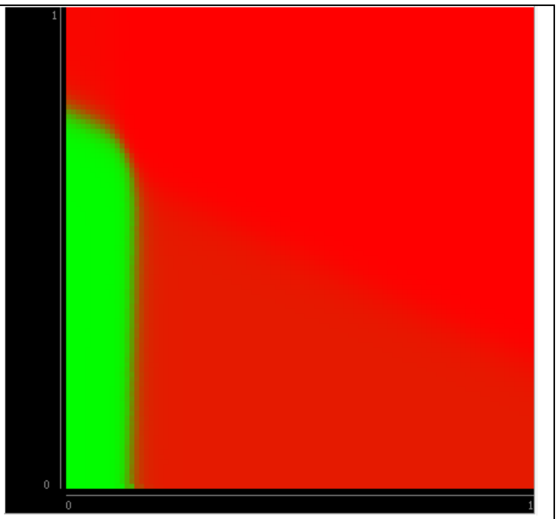

Figure 30. Boundary for Backpropagation for RN and TEXT

Results from Chapter 6 shows KNN, SVM, Decision Tree and Backpropagation have the highest recall and AUC values; however Figure 21, 22, 27 and 28 shows that KNN has noisy decision boundaries which is not suitable for binary classification. Running times of Backpropagation and SVM algorithm are more than other algorithms and it was not possible to implement on full dataset. However Decision Tree shows high recall and AUC values and it is possible to train and test on full dataset in short time which makes the Decision Tree algorithm the most suitable one among other algorithms for binary classification of profile matchings.

APPENDIX B

QUARTILE VALUES FOR MATCHES

Table 15 and 16 shows quartile values of correct and incorrect matches for each feature.

Table 15. Quartile values for correct matches

Feature	**Q1**	**Q2 (Median)**	**Q3**
RN	0.833	1	1
UN	0.519	0.852	1
CN1	0.413	0.601	0.806
CN2	0.480	0.746	0.842
TEXT	0.082	0.111	0.135

Table 16. Quartile values for incorrect matches

Feature	**Q1**	**Q2 (Median)**	**Q3**
RN	0.351	0.45	0.508
UN	0	0.43	0.501
CN1	0	0.413	0.477
CN2	0	0.419	0.482
TEXT	0.007	0.049	0.053