Recognition on Online Social Network by user's writing style

Rodrigo A. Igawa[1], Alex Marino Gonçalves de Almeida[1], Bruno Bogaz Zarpelão[1], Sylvio Barbon Jr[1]

[1]Computer Science Department – State Univerisity of Londrina (UEL)
86057-970 – Paraná – PR – Brazil

alex.marino.almeida@gmail.com, {igawa,brunozarpelao,barbon}@uel.br

Abstract. *Compromising legitimate accounts is the most popular way of disseminating fraudulent content in Online Social Networks (OSN). To address this issue, we propose an approach for recognition of compromised Twitter accounts based on Authorship Verification. Our solution can detect accounts that became compromised by analysing their user writing styles. This way, when an account content does not match its user writing style, we affirm that the account has been compromised, similar to Authorship Verification. Our approach follows the profile-based paradigm and uses N-grams as its kernel. Then, a threshold is found to represent the boundary of an account writing style. Experiments were performed using two subsampled datasets from Twitter. Experimental results showed the developed model is very suitable for compromised recognition of Online Social Networks accounts due to the capacity of recognizing user styles over 95% accuracy for both datasets.*

1. Introduction

Online Social Networks (OSNs) are environments where people discuss and express thoughts and opinions about any subject [Zappavigna 2011]. Currently, OSNs represent a relevant resource of information and research in areas such as Customer Relationship Management (CRM) and Opinion Mining (OM). Knowledge obtained from OSNs such as Twitter and Facebook has shown to be extremely valuable for marketing research companies, public opinion organisations, and other Text Mining purposes [Bahrainian and Dengel 2013, Yu 2012, Zhou et al. 2014, Smailovic et al. 2014, Mostafa 2013, Hsieh et al. 2012]. Since millions of opinions on a certain topic are expressed with simplicity, posting provides rich, easy and unbiased content comprehension [Hassan et al. 2013].

OSNs wide popularity and ease of access have resulted in the misuse of their services. In addition to the privacy preserving issues, OSNs face the challenge of dealing with undesirable users and their malicious activities, spamming for product promotion being one of the most common [Bhat and Abulaish 2013]. To address the problem of malicious activity on social networks, researchers have focused the detection of fake accounts (i.e., automatically created accounts for only spreading malicious content). However, the problem persists once systems that solely detect fake accounts do not

discriminate between fake and compromised accounts. A compromised account is a legitimate account which has been taken over by an attacker to publish fake and harmful content[1,2] . Accounts can be compromised in many different ways, for example, by exploiting a cross-site scripting vulnerability or by using a phishing scam to steal the users credentials. Also, bots have been increasingly used to obtain credentials information for social networking sites on infected hosts [Egele et al. 2013, Grier et al. 2010].

Since fake accounts were mainly created with proposal to cause harm in OSNs, once they are detected, the simplest solution is to delete them. In the meantime, compromised accounts need engaging in a credential recovery process to give back the accounts control to their respective owners [Egele et al. 2013]. Actually, as stated by [Zangerle and Specht 2014], compromised accounts has been the most popular to disseminate fraudulent content. Moreover, a study performed through Twitter revealed that only 16% of the spamming accounts were indeed fake accounts, while the remaining quantity were all compromised accounts [Grier et al. 2010]. The same reality also was seen on Facebook where 97% of malicious accounts were not originally created solely to spamming purpose [Gao et al. 2010].

Considering the scenario described above and also believing that an account behaviour might be recognized by taking into consideration its user writing style. In practice, if some posts are sent in the name of an account and such posts do not present the writing style of its legitimate owner, then we state that the account might have been compromised and malicious contents are being spread. The main limitaiton of such hypothesis is that a considerable amount of text is necessary to extract a user writing style.

Therefore, in this paper, we present a novel study to recognize compromised accounts using only text as resource. Our approach is based on N-grams Authorship Verification (AV) and we focus on recognition of a user based on its writing style. When the writing style of a given user does not match its boundary based on a threshold, then, a warning alarm could be sent out to inform the account owner and malicious posts could be blocked. Also, as seen in [Layton et al. 2010, Uysal and Gunal 2014, Mostafa 2013] text preprocessing, like stopwords removal, can either contribute or disturb text mining tasks, therefore, we also conducted experiments concerning Preprocessing and Corpus size to study their relevance in results.

The remaining of the work is organized as follows: Section 2 presents an overview about compromised accounts and AV along N-grams. In Section 3, details about the proposed approach are described. Section 4 presents the experimental settings

1 http://www.bbc.com/news/world-us-canada-30853311

2 http://www.bbc.com/news/world-us-canada-30785232

to perform our tests along information about both datasets used. Section 5 discusses our results. Section 6 states our conclusion.

2. Related Work

Compromised accounts initially became the object of research interest in e-mail and web services as seen in [Thomas et al. 2011, Khanna and Chaudhry 2012]. In a similar scenario to OSNs, users credentials are stolen using malicious links or phishing techniques [Li et al. 2014, Thomas et al. 2011]. Concerning e-mails, research already conducted work in user levels by using social engineering to emphasize user awareness [Khanna and Chaudhry 2012], while another different approach combined network information, machine learning and content analysis in order to detect harmful content [Thomas et al. 2011].

Some other approaches detected intrusion and compromised accounts in short messages by applying text mining techniques as Authorship Attribution (AA) and AV [Donais et al. 2013, Brocardo et al. 2013, Brocardo et al. 2014]. Their main contribution was to aid the search for cyber criminals [Zhang et al. 2014] or to increase cyber space security and reliability [Donais et al. 2013].

To achieve so, both AA and AV were based on one of two strategies: Stylometry and N-grams. The first one describes text content through attributes which represent writing style markers as lexical, syntactic, content-specific, and idiosyncratic style markers. Lexical attributes are words and character based statistical measures like sentence length. Syntatic attributes include part-of-speech tagger measures. Content-specific attributes are represented by keywords of a given text and idiosyncratic markers are represented by misspellings and grammatical mistakes [Keretna et al. 2013, Ramezani et al. 2013]. N-grams, on the other hand, consist in obtaining frequent co-occurrence patterns in words or character level. A set of most frequent N-grams represents the textual description of a given author, hoping that most frequent patterns would occur more often [Layton et al. 2010, Sun et al. 2010].

Some applications of AA and AV, rather than on e-mails and OSN, include the identification of an author from structureds texts, e.g., textbooks, newspapers, articles, and reviews. In such scenarios, a recent work is found on [Potha and Stamatatos 2014] which performs authorship verification based on N-grams. In such method a given sample of text is assigned to the author in question if the given sample presents a great quantity of N-grams also presented in the author profile. The authors claimed that a great contribution from their work was to use a profile-based paradigm.

Considering AA and AV based on N-grams, which the proposed work is also based on. Some recent work addressed the identification of criminals considering OSN. For example, [Layton et al. 2010] achieved around 70% of accuracy to identify the author of a single tweet within a subset of suspected authors. Such approach was based

in N-grams, and the tweets was assigned to the author presenting the highest quantity of N-grams also presented in the given sample. N values equal to 4, 5, and 6 achieved the best results in their expriments.

Approaches based on Stylometry already performed well to identify an email author. An example of model to achieve such is found on [Iqbal et al. 2013]. The authors experimented many different stylometric features and identified authors by matching the most used stylometric features of each author. In this work, stylometric features are not explored since this work is method based mainly on N-grams. No features from Stylometry were used on this work since dictionaries are necessary to retrieve most of them. Also, Stylometry require different dictionaries and Part of Speech Taggers for different languages.

Regarding the few existent works addressing compromised accounts on OSNs, studies already stated that malicious content are almost completely spread by compromised accounts that were victims of phishing attacks. The detection of malicious accounts is achieved by extracting features from text, webdata and network information to then, classifying it based on machine learning approaches like Random Forest, Support Vector Machines and Logistic Regression [Gao et al. 2010, Stein et al. 2011].

This work is about the proposal of an AV based on N-grams to identify compromised accounts by checking if the writing style of its legitimate user has significantly been changed within a low number of successive posts. As stated earlier, [Layton et al. 2010] performed an AA to assign a given sample to a subset of pre-determined suspects. [Potha and Stamatatos 2014], on the other hand, presented an AV method to identify a single author. However, a work performing an AV method on OSN to address compromised accounts has not been done. The proposed approach might be applied on different kinds of OSN to protect accounts that were compromised by analyzing its legitimate user writing style. This is where the proposed work comes into play.

3. Proposed Approach

The proposed approach is grounded on AV to analyze if an account has been compromised. A compressed version of such approach is also presented in [Igawa et al. 2015].

To represent the legitimate user, it is necessary to extract features from textual content. Such features are obtained using N-grams, as seen in Figure 1.

The main idea behind our proposal is to address compromised accounts problem as a document representation model. By doing so, it would be possible to apply Text Mining tasks to analyze the user writing style.

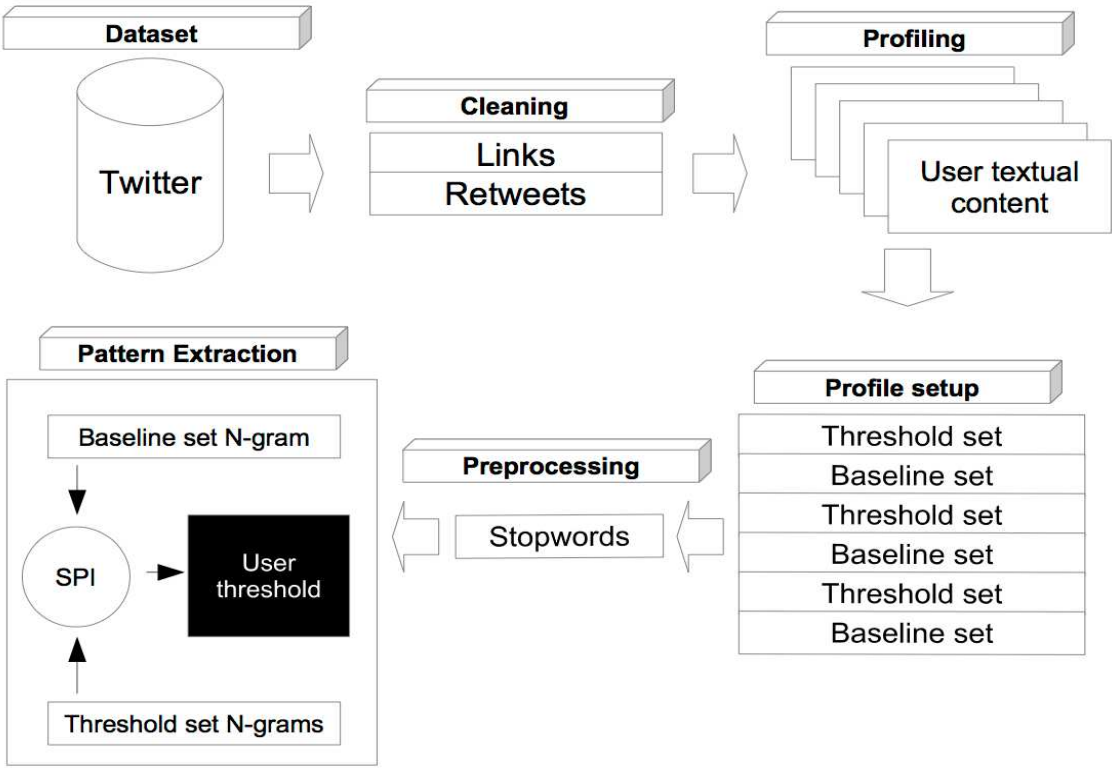

Figure 1. Proposed Approach for User Threshold Estimation

3.1 Dataset

First step is dataset acquisition. Once proposed approach is completely grounded on Text Mining, only text features will be used and, therefore, none additional information beyond the tweets content and their respective authors username are required for our proposal.

3.2 Cleaning

Second step is about Cleaning. Normally, it would be possible to consider any textual content as a part of a document produced by an author. However, as this approach is created to be applied on Twitter users, links and retweets (third part contents) are removed since they do not represent any textual authorship mark.

3.3. Profiling

All remaining text productions are considered authorship samples. Therefore, all contents are concatenated cumulatively following the profile-based paradigm described in [Potha and Stamatatos 2014]. The result of this step, called Profiling, is a document containing all terms written by the user.

3.4. Profile Setup

Then, in Profile Setup, each user is represented as a document whose content is subsampled at the same fraction in two distinct parts: Baseline set and Thresholding set. Each fraction of subsampled document has the same size and is interspersed as shown in Figure 1. This way, subjects discussed by the user during the time will be equally distributed in both sets. This is important to our approach because both sets must have subjects balanced so that the boundary of the writing style can be properly found.

The Baseline set is the text portion which represents user account. This set is used to extract the usual writing style of a user and is kept as one single document as described by the profile-based paradigm in [Potha and Stamatatos 2014]. The Thresholding set is a portion used to find a Simplified Profile Intersection (SPI) threshold to delimitate the user writing style and different from Baseline set, each portion sub- sampled becomes a distinct sample instance. The SPI similarity measure was used in [Potha and Stamatatos 2014, Layton et al. 2010] and is stated to be suitable to different sample sizes. The SPI is calculated as seen in Equation 1, where N_1 and N_2 are two distinct sets of N-grams. Note that SPI is basically a count of N-grams that exist in both sets without considering frequency.

$$SPI(N_1,N_2) = [N_1 \cup N_2[\quad (1)$$

3.5. Preprocesing

After Profile Setup process, Preprocessing techniques (in this approach, stopwords removal) can be performed to improve the effectiveness of our approach. In this work, we explore some combination of Preprocessing concerning precision and accuracy to recognize accounts textual content.

3.6. Writing Style Extraction

To obtain the SPI threshold in Writing Style Extraction step, most frequent N-grams are extracted from Baseline set and most frequent N-grams are also extracted from each fraction in Thresholding set. Table 1 shows an example of the 10 most frequent N-grams extracted from a random user found in one our datasets. In such example the N used is 4, “and_” was the most occurrent pattern, found 202 times within its user posts, 200 occurrencies for “the_” and so on. The symbol “_” represents white space occurrence.

In this work, all grams, and not only the 10 most frequent, were considered to perform experiments.

Table 1. Example of 10 most frequent N-grams (N=4) from a random user

Frequency	Gram
202	“and_”
200	“the_”
109	“ciat“
108	“ecia”
108	“iate”
107	“ppre”
107	“prec”
107	“reci”
104	“Appr”
95	“S___”

Then, SPI is used to calculate similarity between Baseline set and each sample of Thresholding set N-grams, generating a vector of SPI distances, SPI_{vector}. Such vector is used to obtain a SPI threshold and details towards formulas used to obtain SPI threshold are shown in Section 4. Any future portion of text posted in by this account that presents SPI measure lesser than threshold is considered an intrusion and the account is considered as compromised

4. Methodology

Twitter, the OSN used in this work, is known as a micro blogging service. Unlike other OSNs, Twitter is known by short posts (140 characters at maximum) users do to express thoughts, opinions and feelings [Zappavigna 2011]. These short texts, named tweets, are available publicly as default, and are immediately broadcasted to the users followers [Bliss et al. 2012].

The Twitter Developer Team offers a streaming service that delivers other developers low latency access to Twitter’s global stream of data. The tweets sets from both dataset used as samples in our experiments were collected by [Yang and Leskovec 2011] and [Li et al. 2012] using this service.

In experiments3, only a small subsample from original datasets were used. Details concerning both datasets used in experiments are shown in Section 4.1 and Section 4.2. Considering Dataset I, only a small subsample from original Dataset I was used. Addressing Dataset II, also a subsample was used, however, this time, a substantially larger set of tweets were used.

Then, all tweets were grouped by authors username cumulatively following the profiling step described in Section 3. As specified in our approach, links and retweets were removed since they do not infer any information about users writing pattern. All remaining textual content was included in our tests.

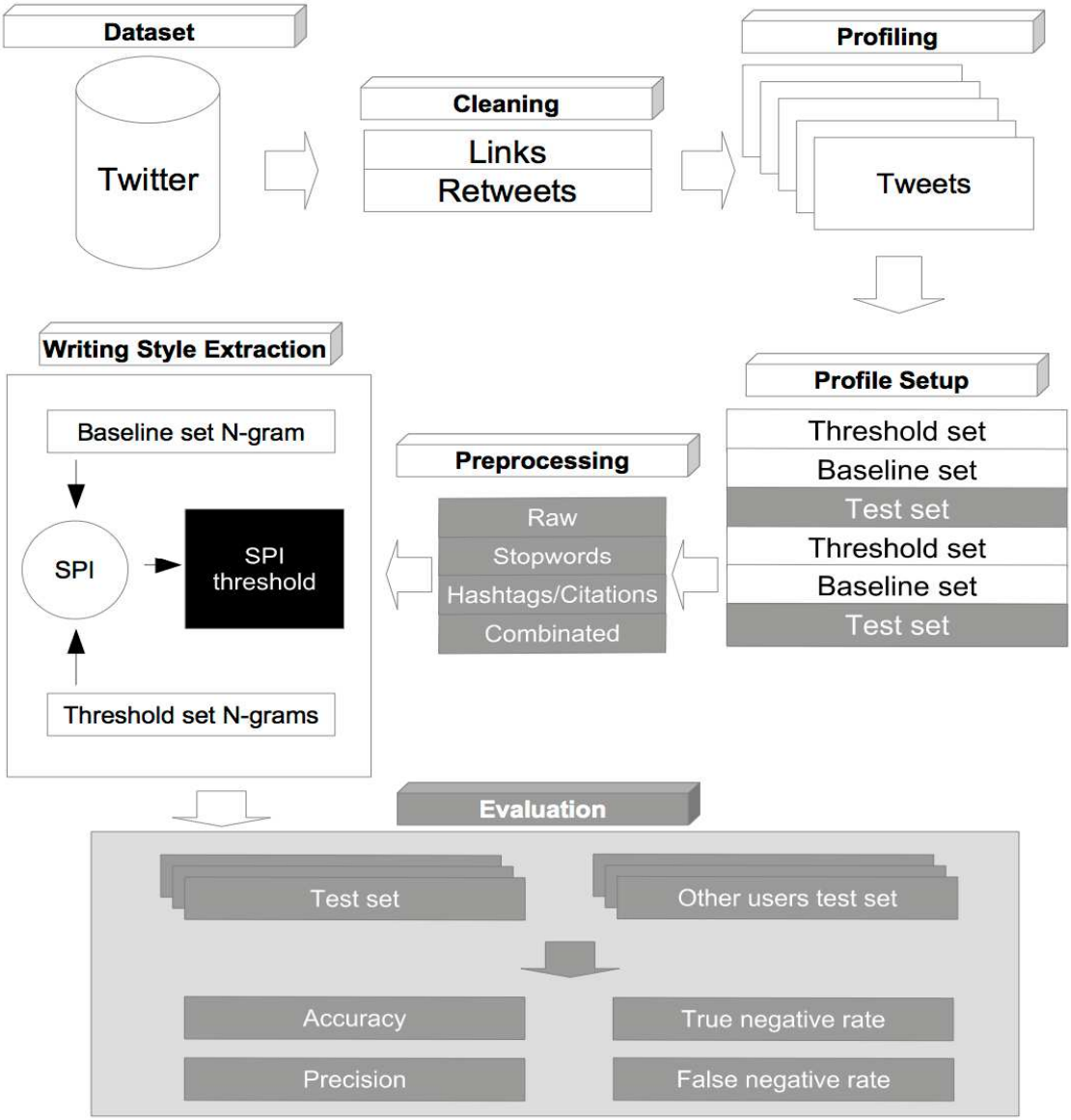

Figure 2: Experimental Settings Overview

3 https://wiki.cites.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012

In Figure 2, gray parts represent the experimental settings. The Profile Setup process was performed to separate textual contents in 3 distinct parts: Baseline set, Thresholding set (as proposed in Section 3) and Test set. The last set is used to evaluate our method's efficiency and is a representation of future portions of text posted. The user's Test set is completely used along randomly selected Test set instances from other users to check how adequate is the obtained threshold. Our intention to use other users test set among the own user test set is to simulate a situation where the legitimate account has been compromised and harmful posts are written. In such cases, other users test set represent posts from invaders, therefore, it is desirable to obtain SPI measures in instances from other users test set lesser than threshold, while own user test set are intended to present SPI greater than threshold.

Another concerning about the Profile setup is to study the size of each splitted part. This is considered an important issue of this work once the size used presenting better results would be the amount of words necessary to recognize accounts textual contents. In our experiments, were used 11 different sizes ranging from 50 to 100 words.

Also, concerning Preprocessing, 4 tests concerning their influence were performed: a) Raw (i.e, no preprocessing), b) Hashtags and Citations removal, c) Stopwords removal and d) Combined preprocessing (i.e, Hashtags, Citations and Stopwords removal). The idea behind these tests is to study the influence of disposable terms concerning precision and accuracy to recognize account textual content.

One last issue addressing experimental setting is which values of N to be used on N-grams. These values are applied including the Corpus size and preprocessing settings. To decide such, we used the results of AA reported from [Layton et al. 2010], and therefore, the values used are 4, 5, and 6.

Therefore, the complete experimental setting consists in 132 experiments, for each SPI threshold measure shown in Table 2, for each dataset. Such 132 different configurations cover our 3 different N-grams values (4, 5 and 6), 11 combinations towards Corpus size and the 4 combination dealing 2 preprocessing techniques4.

Table 2. Measures used to evaluate recognition rate

Name	Formula
Threshold I - Minimal SPI	min(SPIvector)
Threshold II - Augmented Minimal SPI	min(SPIvector) + std(SPIvector)
Threshold III - Decreased Minimal SPI	min(SPIvector) − std(SPIvector)

4 Implementation made and tests performed on MacBook Pro (13-inch, Mid 2012), Processor: 2.5 GHz Intel Core i5, Memory: 4 GB 1600 MHz DDR3

Threshold IV - Augmented Averaged SPI	avg(SPIvector) + std(SPIvector)
Threshold V - Decreased Averaged SPI	avg(SPIvector) − std(SPIvector)

Aiming to keep experiments always balanced to enable comparisons to each other, we defined that each Test set and Threshold set were composed by 10 instances of same size (ranging from 50 to 100 words). In the Evaluation step, the user being recognized always used its entire Test set (i.e, 10 instances of text) along 10 instances randomly selected from other users Test sets. As stated previously, other users Test sets represent invaders.

To evaluate our method efficiency, we used 4 well known statistical measures found in [Olson and Delen 2008] and their equations are shown in Table 3 where TP are user test set instances presenting SPI greater than threshold. TN are other users test set presenting SPI lesser than threshold, FN are user test instances presenting SPI lesser than threshold and FP are other users test instances presenting SPI greater than threshold. Analysis results and discussion towards all experiments are presented in Section 5.

Table 3. Measures used to evaluate recognition rate

Name	Equation
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
False Negative Rate	$\frac{FN}{FN+TP}$
True Negative Rate	$\frac{TN}{TN+FP}$

In practice, TP are instances from the account correctly recognized. TN are instances that are not from the account that are correctly identified as not from the account, and thus, a possible invader. FN are instances from the account that were not recognized as from the account user. In this case, the user made a post but our method did not recognize him. FP are instances that are not from the account but were recognized as from the account. This time, someone other than its legitimate owner would have posted and our method did perceive.

4.1. Dataset I

Dataset I is a subsample of the dataset used in [Yang and Leskovec 2011]. On its original form, the dataset corresponds to 467 million Twitter posts from 20 million users

covering a 7 months period in 2009. Authors claimed that this dataset corresponded to 30% of overall tweets of that time. For every single tweet, there are three information available: author, time and content. However, unlike Dataset II, it is necessary to use a parser to obtain the tweets of different users separately.

Considering our experiments, 50 user were randomly selected. Such a small dataset was considered to evaluate the proposed method in a critical scenario were a new OSN is born and only a few users are present. Even considering such a small group the proposed method needs to distinguish users correctly. A second reason to use this size of dataset is individual analysis. Taking more users into consideration would not enable analysis as seen in Figure 4.

Dataset I, in a very similar way from Dataset II, is not collected using one or more topics. Instead of using Twitter API as most of users do, the authors from both works where Dataset I and Dataset II are found collected tweets from users and not from query. This way, a dataset without specifics topics are possible.

4.2. Dataset II

Dataset II is a subsample of the dataset used in [Li et al. 2012]. On its original form, the dataset contains 284 million following relationships, 3 million users profiles and over 50 million tweets. The crawling was performed by the authors during May 2011. A very interesting point concerning such dataset is that almost every single user crawled has around 500 tweets which supplies suitable amounts of text to be used.

Still, concerning the text availability in such dataset, once obtained, tweets belonging to one specific accounts are already separated in different files named by the user number ID. In our experiments, 250 were randomly selected to be used in experiments. As already informed, most of users have around 500 tweets, so no problems concerning a random selection would be found. One might question if 250 users are enough to evaluate the developed model. Actually, no loss in accuracy is found by changing the size of dataset used. In practice the the proposed approach showed to be invariant considering dataset size, as seen in Section 5.

5. Results and Discussion

As described previously, 132 experiments were performed concerning all possible combinations within N = 4, 5, 6; Corpus size in each splitted portion ranged from 50 to 100 words and 4 combinations of text preprocessing. These 132 experiments were performed once for each threshold measure, in both datasets. Totally, 1320 experiments were performed.

Our first discussion is about threshold measure. Table 2 shows all five different measures used in experiments to obtain an account SPI threshold. The results about such measures is shown in Figure 3 where blue bars represent results concerning Dataset I

and yellows bars represent results concerning Dataset II. Both bars illustrate performance of thresholds in terms of accuracy.

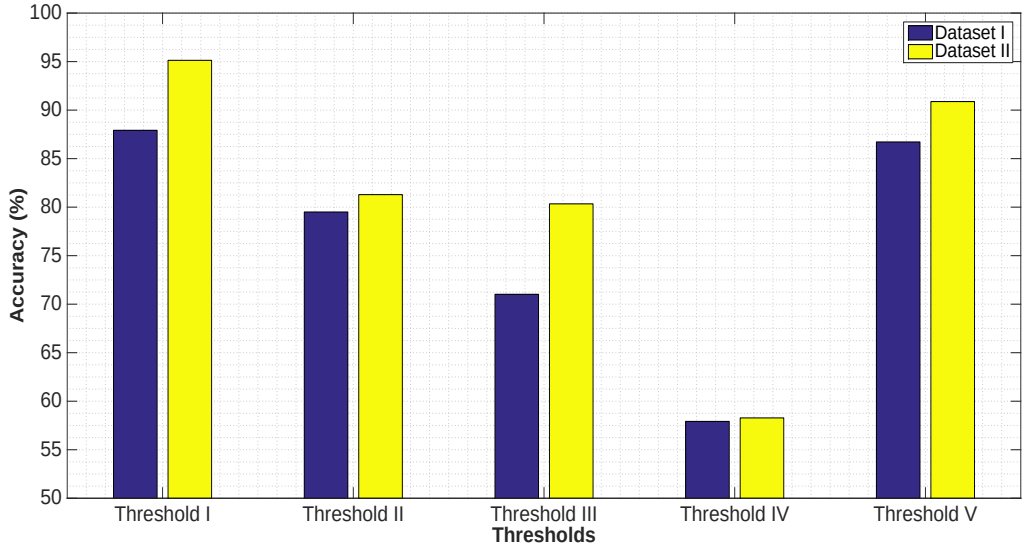

Figure 3. Geral results towards thresholds accuracy

As it is possible to realize, most successful results are from threshold I and threshold V. Others thresholds, as threshold IV, presented a lower result in accuracy since its value is represented by a higher value of SPI measure. In practice, the higher the SPI threshold is, the higher its precision will be. However, false negative rate also increases along higher values of SPI thresholds which decreases its final accuracy. This is specifically seen for threshold IV (highest SPI threshold of all) on both Dataset I and II. Details concerning such statistics for both dataset and all five experimented thresholds are presented on Table 4.

Table 4. Threshold measures and detailed results

	Precision	Accuracy	TNR	FNR	Threshold
Dataset I	83.35%	85.63%	75.48%	4.22%	I
Dataset I	94.08%	79.50%	94.53%	35.53%	II
Dataset I	67.34%	71.06%	42.12%	0.00%	III
Dataset I	97.23%	57.92%	99.61%	83.76%	IV
Dataset I	90.38%	86.72%	88.64%	15.20%	V
Dataset II	93.11%	95.13%	90.27%	0.00%	I
Dataset II	98.70%	81.29%	98.82%	36.25%	II
Dataset II	77.10%	80.34%	60.67%	0.00	III
Dataset II	97.15%	58.28%	99.93%	83.37%	IV
Dataset II	96.92%	90.88%	96.48%	14.72%	V

Still concerning Table 4, in order to consider a balance from higher precisions and accuracies along false negative rates, the most successful threshold in experiments is threshold I. On dataset I, threshold I (85.63%) did not obtained accuracy higher than threshold V (86.72%), however, it obtained more appropriate rates of false negative, 4.22% for threshold I and 15.20% for threshold V.

Thus, further discussions about specific results concerning Dataset I and Dataset II take into consideration only results obtained by using Threshold I.

5.1. Results on Dataset I

Discussion concerning Dataset I takes into consideration only threshold I as its SPI threshold. Results considering both accuracy and precision are shown in Table 5 and 6, highest and lowest results in accuracy respectively, without considering other layers as N or Corpus Size.

It is notable that the top settings achieved excellent results, ranging from 94.10% to 95.80% accuracy (i,e. correctly classified instances) and also presented excellent results in terms of true negative rate ranging from 88.40% to 91.60% which indicates that our method is capable of infer when the content does not correspond to its legitimate user writing pattern.

Another important issue to be observed is the performance for different combinations of preprocessing in both Table 5 and Table 6. All 10 top results achieved their accuracies without removing hashtags and citations. On the other hand, all 10 least accurate experiments applied hashtags and citations removal achieving poor results. Our first conclusion by overviewing the experiments is that hashtags and citations carry information about the writing style of a user textual content, once they indicate subjects discussed and people frequently contacted.

Table 5. Top results in accuracy

N	C. Size	Prec	Acc	TNR	FNR	Hashtags/Cit	Stopwords
6	100	93.97%	95.80%	91.60%	0.00%	Not removed	Removed
5	100	93.43%	95.70%	91.40%	0.00%	Not removed	Removed
6	95	93.59%	95.50%	91.00%	0.00%	Not removed	Removed
6	90	93.64%	95.30%	90.80%	0.20%	Not removed	Removed
5	90	93.36%	95.10%	90.40%	0.02%	Not removed	Removed
6	70	92.72%	95.00%	90.20%	0.02%	Not removed	Removed
4	100	92.28%	94.60%	89.20%	0.00%	Not removed	Removed
5	70	92.45%	94.60%	89.40%	0.02%	Not removed	Removed
6	85	92.57%	94.20%	89.00%	0.06%	Not removed	Removed
5	95	91.39%	94.10%	88.40%	0.02%	Not removed	Removed

Table 6. Lowest results in accuracy

N	C. Size	Prec	Acc	TNR	FNR	Hashtags/Cit	Stopwords
6	55	72.31%	77.60%	55.40%	0.20%	Removed	Removed
6	65	73.75%	77.40%	61.20%	6.40%	Removed	Not removed
6	50	75.08%	77.20%	63.20%	8.80%	Removed	Not removed
4	60	75.47%	77.10%	64.40%	10.20%	Removed	Not removed
4	70	74.53%	76.70%	61.80%	8.40%	Removed	Not removed
4	55	71.73%	76.60%	53.80%	0.60%	Removed	Removed
5	50	74.03%	76.10%	60.60%	8.40%	Removed	Not removed
4	50	73.89%	75.60%	60.20%	9.00%	Removed	Not removed
4	55	71.30%	74.70%	55.80%	6.40%	Removed	Not removed
6	55	70.97%	73.80%	55.40%	7.80%	Removed	Not removed
5	55	70.62%	73.30%	55.00%	8.40%	Removed	Not removed

Still concerning the preprocessing issue, a detailed result from the top 1 experimental setting in Table 5 using Corpus size = 100 and N = 6 is shown in Table 7 in terms of accuracy. Just by removing hashtags and citations, a loss in accuracy is found, falling from 91.90% to 86.10% accuracy. By removing only stopwords it is still possible to increase 5.0% accuracy. This implies that pronouns, articles and prepositions do not help to recognize a user writing style using our approach. One last observation about preprocessing is: a combination of hashtags/citations and stopwords removal achieves the lowest results of the 4 combinations once it uses only a little part of writing not including stopwords, hashtags and citations.

A discussion about the top 1 setting in Table 5 is illustrated by Figure 4 and shows accuracy considering each user. The setting achieved 100% of correctly recognized in many cases, however, to accounts number 4, 15, 25 and 27, accuracy below 80% were achieved. These users presented a very unstable writing style using a high quantity of prepositions and almost nothing of jargons and emoticons, making their writing difficult to distinguish. In all other cases, the setting obtained satisfactory results.

Corpus size influence on our approach is illustrated by Figure 5. Before any discussion about this view, it is necessary to observe that the Corpus size is not used only to split in Profile Setup process, but also implies in the number of words necessary to perform proposed approach with satisfactory results.

Considering so, the fact that none setting size used in our experiments presented outliers and also achieved balanced quartiles is encouraging. It implies that our method has a stable range of accuracy independently of the amount of text used. A descending gradient observed on accuracy using 100 to 50 words is justifiable once less words also means less N-grams to be extracted and possibly less accuracy. Therefore, the box plot

states that the most considerable size to be used in our dataset is 100 words while the most inappropriate is 50.

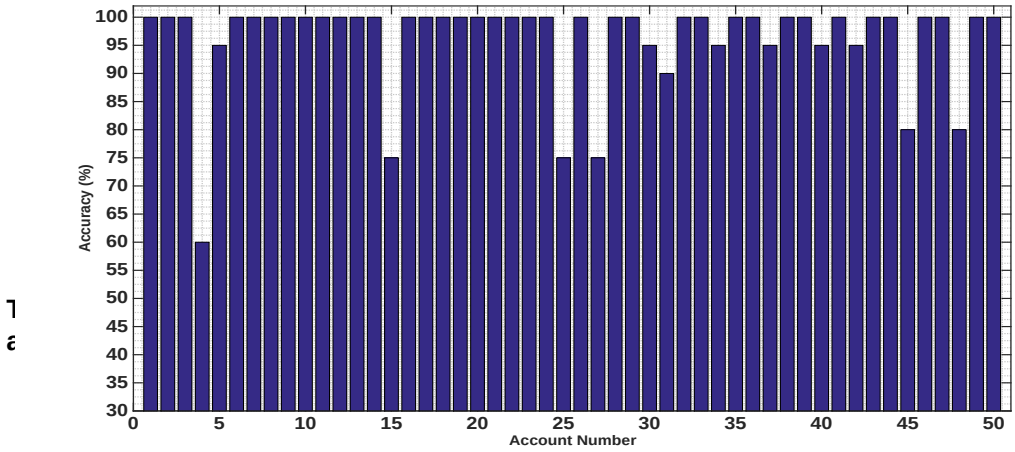

Figure 4. Individual accuracy on recognition of compromised

One last consideration about our approach performance on Dataset I is the thresh- old value and its relation to each user writing style. Most part of users obtained 100% accuracy as shown in Figure 4. These users are represented in Figure 6, as Case III, where the obtained threshold is suitable to separate writing styles from the legitimate user and other users. Case I and II represent users that by presenting too many stopwords as part of their writing styles and using a small quantity of emoticons, jargons, hashtags or citations obtained a threshold value unable to correctly separate writing styles and, there- fore, obtained a significant number of false negative (i.e, writing style from different users being recognized as the user in question).

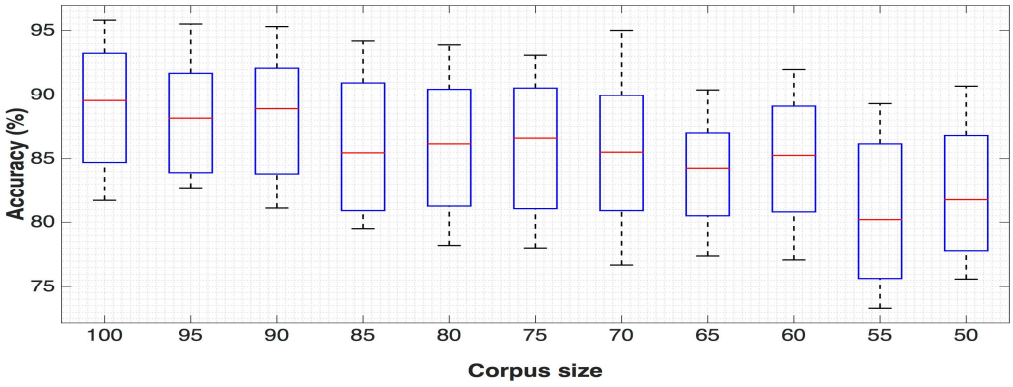

Figure 5. The influence of Corpus size on baseline accuracy

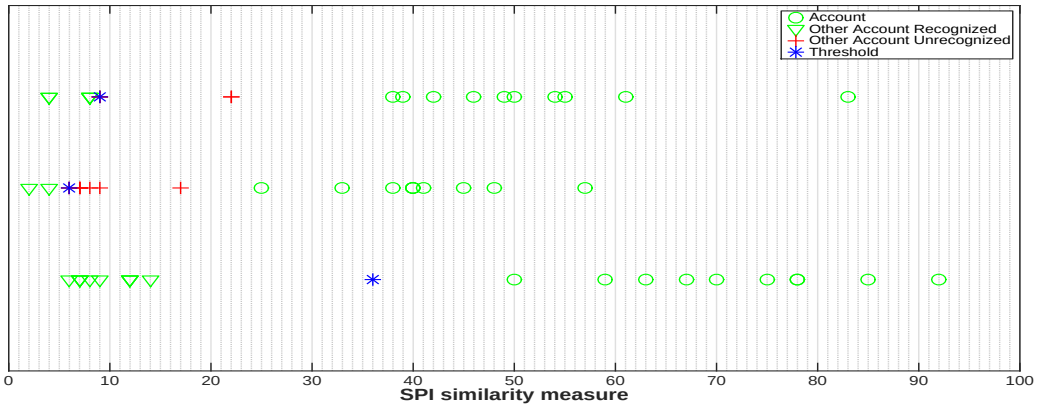

Figure 6. Threshold testing for compromised accounts

5.2. Results on Dataset II

As specified on Section 4.2, Dataset II consists on a collection of 250 accounts. This way, a detailed result concerning individual accuracies for each accounts as shown in Figure 4 about all 50 accounts on Dataset I would not be visible. However, as considered on previous discussions towards Dataset I, our discussions addressing Dataset II also take into consideration only results obtained by using threshold I as its SPI threshold.

This way, Table 8 and Table 9 present highest and lowest results in accuracy respectively. By overviewing experiments on Dataset II, we note that, although a larger collection of accounts was used, good results in terms of accuracy were obtained. Top results showed on Table 8 ranged from 98.50% to 97.64% accuracy, while lowest results ranged from 92.26% to 90.36% accuracy.

Our first issue towards both tables is also the same issue considered for Dataset I: combination of preprocessing tasks. As it can be seen on Table 9, all the lowest results in accuracy removed hashtags and citations. On the other hand, all the top 10 results in terms of accuracy shown in Table 8 did not removed neither hashtags or citations. Table 10 shows results about the combination of preprocessing tasks concerning the top results in terms of accuracy from Table 8. Just by removing hashtags, a loss around 2.0% accuracy is found. By removing hashtags and citations only, an increase of 0.12% is achieved. This is also a similar result from Dataset I.

Table 8. Top results in accuracy on Dataset II

N	C.Size	Prec	Acc	TNR	FNR	Hashtags/Cit.	Stopwords
6	100	97.62%	98.50%	97.00%	0.00%	Not removed	Removed
6	100	97.33%	98.38%	96.76%	0.00%	Not removed	Not removed
5	100	97.30%	98.28%	96.56%	0.00%	Not removed	Removed
6	90	97.20%	98.26%	96.52%	0.00%	Not removed	Not removed
6	90	97.23%	98.24%	96.48%	0.00%	Not removed	Removed
6	80	96.77%	97.84%	95.68%	0.00%	Not removed	Not removed
5	90	96.44%	97.78%	95.56%	0.00%	Not removed	Not removed
6	95	96.61%	97.76%	95.52%	0.00%	Not removed	Not removed
5	100	96.39%	97.68%	95.36%	0.00%	Not removed	Not removed
5	80	96.48%	97.64%	95.28%	0.00%	Not removed	Removed

Table 9. Accuracy lowest results on Dataset II

N	C.Size	Prec	Acc	TNR	FNR	Hashtags/Cit.	Stopwords
4	70	89.32%	92.26%	84.52%	0.00%	Removed	Not removed
4	65	89.09%	92.06%	84.12%	0.00%	Removed	Removed
4	60	89.31%	92.00%	84.00%	0.00%	Removed	Removed
4	50	88.61%	91.86%	83.72%	0.00%	Removed	Not removed
4	80	88.87%	91.76%	83.52%	0.00%	Removed	Not removed
4	65	88.74%	91.74%	83.48%	0.00%	Removed	Not removed
4	55	88.41%	91.56%	83.12%	0.00%	Removed	Not removed
5	50	88.78%	91.54%	83.08%	0.00%	Removed	Removed
4	55	88.58%	91.52%	83.04%	0.00%	Removed	Removed
4	50	87.49%	90.36%	80.72%	0.00%	Removed	Removed

Table 10. The influence of text Preprocessing techniques on compromised accounts recognition on Dataset II

Preprocessing	Mean	Standard Deviation
Raw	98.38%	8.15%
Hashtags/Citations Removal	96.96%	8.80%
Stopwords Removal	98.50%	4.02%
Combinated Preproc.	96.58%	4.35%

Just as stated about Dataset I, on Dataset II, which is a completely different datasets and also consists of a larger collections of accounts, it is possible to realize that hashtags and citations carry information about the writing style of a user textual content, once they indicate subjects discussed and people frequently contacted. Such elements show very important marks of writing style to be used on authorship on both datasets.

Corpus size influence on our approach considering Dataset II is illustrated by Figure 7. Any discussion about this issue has to also take into consideration that Corpus Size influence on our approach also implies in the number of words that our model

would require if implemented on a real scenario. Choosing a Corpus Size of 100 words means that the split step on profile process used 100 words in each sample, but also means that in a real scenario, it takes 100 words to our model perform a prediction about an account writing style. Considering so, we state the only amount of Corpus Size that could present results not stable on real scenario would be 75. On Figure 7, 75 words is the only amount which presented an outlier. All other amounts of words experimented achieved stable results not only by not presenting outliers, but also by presenting satisfactory box sizes. Just as on Dataset I, a descending gradient can be observed on accuracy ranging from 100 to 50 words. It is justifiable once less words also means less N-grams to be extracted in order to delimit an account writing style. Thus, as also stated about the corpus size case on Dataset I, the 100 words is the amount of words which achieved the highest results, while 50 words achieved the lowest results, both in maximum and minimum accuracy.

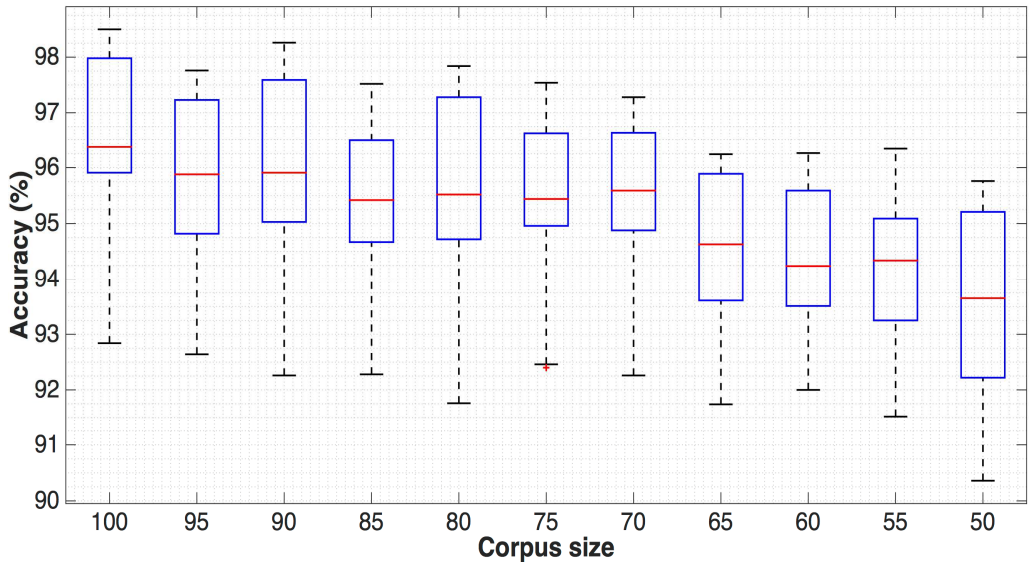

Figure 7. The influence of Corpus size on baseline accuracy on dataset II

In summary, the results presented above are very encouraging concerning compromised accounts on OSNs. In both datasets, different measures as precision, accuracy, true negative rate and false negative rate were used in order to evaluate the proposed approach.

Regarding preprocessing, results showed that hashtags and citations should not be removed once they represent people and subjects usually talked about. On the other hand, stopwords removal contributed on better results. Considering corpus size and N-grams, results showed that 100 words was most desirable amount along N=6 on both datasets. Such setting achieves the highest accuracy over all other settings tested.

6. Conclusion and Future Work

Compromised accounts represent a subset of the malicious accounts which deleting should not be an option. Once detected, compromised accounts need to engage in a credentials recovery process to give back the accounts control to their respective owners. Current works about malicious accounts rely on features from text, webdata and network information to classify an account. Also, no related work was performed considering each user individually. The most similar classification problem can not be directly compared because addressed different user behaviours after compromised like moving out to a new account or changing credentials.

One advantage from our approach is that only text is used as resource once it is grounded on Text Mining. Although it was tested on Twitter in our experiments, our developed method is applicable on any OSN. Also, due to the fact that this work is the first to depend only on text to recognize compromised accounts, our approach concerned about the Corpus size necessary to recognize compromised accounts, desirable preprocessing to obtain better results and which N use in N-grams calculation to improve the approach results.

In this work, one important consideration is that warning systems should not recognize a legitimate user as an invader. So, we design a model for avoiding false positive of compromised accounts. Thus, we studied and proved that the top experimental setting presented on both datasets. An important contribution was that on a corpus size of 100 words and stopwords removal as the only text preprocessing. Using 6-grams, it was achieved good results of precision and accuracy along few occurrences of false negative. This fact implies that our method would rarely claim compromised accounts by textual content when actually it was not compromised.

Another very important issue to be addressed is that, in this work, a AV based on N-grams was presented. A work considering stylometric approaches, e.g., considering the number of prepositions, articles, and pronouns would achieve satisfactory results. However, in this proposed approach based on N-grams, removing stopwords in necessary to achieve higher accuracies.

Considering short texts scenario of Twitter, a use of 100 words are an acceptable amount as validated by the results. Our experiments had a mean of 14.6 words per tweet. In practice, it is possible to recognize a user account by textual content based on 6-10 tweets with 95% accuracy with 91% true negative.

For future work it would be of great interest to study a method dealing only with those cases of low accuracy. Another relevant issue to be treated in future works is the amount of text used to extract a user writing pattern.This way, our method’s accuracy could be increased. This study could experiment other N-grams measurements focused on special cases of authors.

References

Bahrainian, S.-A. and Dengel, A. (2013). Sentiment analysis and summarization of twitter data. In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on, pages 227–234. IEEE.

Bhat, S. Y. and Abulaish, M. (2013). Community-based features for identifying spammers in online social networks. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 100–107. ACM.

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., and Dodds, P. S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. Jour- nal of Computational Science, 3(5):388–397.

Brocardo, M. L., Traore, I., Saad, S., and Woungang, I. (2013). Authorship verification for short messages using stylometry. In Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on, pages 1–6. IEEE.

Brocardo, M. L., Traore, I., and Woungang, I. (2014). Authorship verification of e-mail and tweet messages applied for continuous authentication. Journal of Computer and System Sciences, pages –.

Donais, J. A., Frost, R. A., Peelar, S. M., and Roddy, R. A. (2013). Summary: A system for the automated author attribution of text and instant messages. In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, pages 1484–1485. IEEE.

Egele, M., Stringhini, G., Kruegel, C., and Vigna, G. (2013). Compa: Detecting compromised accounts on social networks. In NDSS.

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pages 35–47. ACM.

Grier, C., Thomas, K., Paxson, V., and Zhang, M. (2010). @ spam: the underground on 140 characters or less. In Proceedings of the 17th ACM conference on Computer and communications security, pages 27–37. ACM.

Hassan, A., Abbasi, A., and Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. In Social Computing (SocialCom), 2013 International Conference on, pages 357–364. IEEE.

Hsieh, L.-C., Lee, C.-W., Chiu, T.-H., and Hsu, W. (2012). Live semantic sport highlight detection based on analyzing tweets of twitter. In Multimedia and Expo (ICME), 2012 IEEE International Conference on, pages 949–954. IEEE.

Igawa, R. A., de Almeida, A. M. G., Zarpelao, B. B., and Barbon, Jr, S. (2015). Recognition of compromised accounts on twitter. In Proceedings of the Annual Conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective - Volume 1, SBSI 2015, pages 2:9–2:14, Porto Alegre, Brazil, Brazil. Brazilian Computer Society.

Iqbal, F., Binsalleeh, H., Fung, B. C., and Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. Information Sciences, 231:98–112.

Keretna, S., Hossny, A., and Creighton, D. (2013). Recognising user identity in twitter social networks via text mining. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, pages 3079–3082. IEEE.Khanna, S. and Chaudhry, H. (2012). Anatomy of compromising email accounts. In Information and Automation (ICIA), 2012 International Conference on, pages 640– 645.

Layton, R., Watters, P., and Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. In Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second, pages 1–8. IEEE.

Li, C.-H., Hsu, F.-H., Chen, S.-J., Wang, C.-S., Chen, Y.-H., and Hwang, Y.-L. (2014). Hawkeye: Finding spamming accounts. In Network Operations and Management Sym- posium (APNOMS), 2014 16th Asia-Pacific, pages 1–4. IEEE.

Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. In KDD, pages 1023–1031.

Mostafa, M. M. (2013). More than words: Social networks text mining for consumer brand sentiments. Expert Systems with Applications, 40(10):4241–4251.

Olson, D. L. and Delen, D. (2008). Advanced data mining techniques. Springer Science & Business Media.

Potha, N. and Stamatatos, E. (2014). A profile-based method for authorship verification. In Artificial Intelligence: Methods and Applications, pages 313–326. Springer.

Ramezani, R., Sheydaei, N., and Kahani, M. (2013). Evaluating the effects of textual features on authorship attribution accuracy. In Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on, pages 108–113. IEEE.

Smailovic, J., Grcar, M., Lavrac, N., and Znidarsic, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. Information Sciences.

Stein, T., Chen, E., and Mangla, K. (2011). Facebook immune system. In Proceedings of the 4th Workshop on Social Network Systems, page 8. ACM.

Sun, J., Yang, Z., Wang, P., and Liu, S. (2010). Variable length character n-gram ap- proach for online writeprint identification. In Multimedia Information Networking and Security (MINES), 2010 International Conference on, pages 486–490. IEEE.

Thomas, K., Grier, C., Ma, J., Paxson, V., and Song, D. (2011). Design and evaluation of a real-time url spam filtering service. In Security and Privacy (SP), 2011 IEEE Symposium on, pages 447–462. IEEE.

Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1):104–112.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In Pro- ceedings of the fourth ACM international conference on Web search and data mining, pages 177–186. ACM.

Yu, S. J. (2012). The dynamic competitive recommendation algorithm in social network services. Information Sciences, 187:1–14.

Zangerle, E. and Specht, G. (2014). Sorry, i was hacked: a classification of compromised twitter accounts. In Proceedings of the 29th Annual ACM Symposium on Applied Computing, pages 587–593. ACM.

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on twitter. New Media & Society, 13(5):788–806.

Zhang, C., Wu, X., Niu, Z., and Ding, W. (2014). Authorship identification from unstructured texts. Knowledge-Based Systems.

Zhou, X., Wu, S., Chen, C., Chen, G., and Ying, S. (2014). Real-time recommendation for microblogs. Information Sciences, 279:301–325.