Stylometric Linkability of Tweets

Mishari Almishari
King Saud University
mialmishari@ksu.edu.sa

Mohamed Ali Kaafar
NICTA
dali.kaafar@nicta.com.au

Ekin Oguz
Gene Tsudik
University of California, Irvine
{eoguz,gene.tsudik}@uci.edu

ABSTRACT

Microblogging is a very popular Internet activity that informs and entertains a large number of people via terse messages; e.g., tweets on Twitter. Even though microblogging does not emphasize privacy, authors can easily hide behind pseudonyms and multiple accounts on the same, or across multiple, site(s). In this paper, we explore stylometric linkability of tweets. Our results clearly demonstrate that multiple sets of tweets by the same author are easily linkable even when the number of possible authors is large. This is also confirmed by showing that linkability holds for a set of *actual Twitter users* who admittedly tweet via multiple accounts.

Note: a full version of this paper is available in [7].

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: Privacy

Keywords

Linkability; Microblogging; Author Attribution

1. INTRODUCTION

Microblogging offers fast and highly scalable information sharing, allowing multitudes of users to disseminate pithy messages to news-hungry and attention-challenged followers or subscribers. For social or professional purposes, users share their thoughts, interests and sometimes express highly sensitive or controversial opinions. Twitter, the most prominent microblogging site, has grown to a truly global service with a very large number of users; both tweeters and followers.

As its popularity grew, microblogging has become a rich source of information about individuals. Although it is hard to argue that the existence of public messages violates privacy of their authors – since they are the one who makes their utterances public in the first place – the potential to de-anonymize multiple user accounts and to link messages (or sets thereof) produced by the same author is a threat to privacy. This is primarily because multiple accounts owners often expect each set of messages to remain within the boundaries of the originating account.

On the other hand, microblogging site operators and law-enforcement agencies might benefit from techniques that link accounts within a site or across multiple sites. Legitimate reasons might include: (1) identifying spammers and phishers who hide behind multiple accounts to evade detection, (2) tracking or correlating messages pertaining to illegal activities, such as terrorism incitement, pedophilia or human/drug trafficking.

The goal of this work is to assess linkability of tweets by exploring stylometric similarities between multiple sets of tweets by the same author. We analyze two large datasets, each containing over $8,000$ Twitter accounts. Using simple probabilistic models (Naïve Bayes) and simple textual features, we show that tweets are highly linkable even if the overall number of tweeters is very large; e.g., $> 8,000$. By analyzing simple letter frequencies, we can – in some scenarios – link sets of tweets with $> 90\%$ accuracy. We also extend this analysis to demonstrate linkability of sets of tweets of actual users who admittedly operate multiple Twitter accounts. A complete full-length version of this paper can be found in [7].

2. BACKGROUND: NB MODEL

Naïve Bayes (NB) [8] is a probabilistic model based on the so-called Naïve Bayes assumption, which states that all features/tokens are conditionally independent, given some category. In our case, the category is a tweeter (user). Given a document with a set of tokens/features: $token_1, token_2, \ldots, token_n$, NB model computes the corresponding user model as: $User = argmax_U P(U|token_1, \ldots, token_n)$ where U varies over all distinct users in a dataset, in order to maximize the probability $P(U|token_1, \ldots, token_n)$ of categorizing a document as belonging to a specific user.

Using the Bayes Rule[8], Naïve Bayes and the assumption that $P(U)$ is the same for all users, finding a matching $User$ for $token_1, token_2, \ldots, token_n$ for

WPES'14, November 03 2014, Scottsdale, AZ, USA.

ACM ACM 978-1-4503-3148-7/14/11 ...$15.00.
http://dx.doi.org/10.1145/2665943.2665966 .

a given tweeter profile boils down to: $User = argmax_U P(token_1|U) \cdots P(token_n|U)$

To avoid the under-flow problem, we consider *log* of the products, which results in [1]: $User = argmax_U \sum_i logP(token_i|U)$

3. DATASET AND SETTINGS

Dataset. We use a portion of the very large dataset crawled by Yang et al. [12], that spans an approximately six-month period from June to December, 2009 [2]. It contains $> 4 * 10^8$ tweets, authored by $> 15 * 10^6$ million users, of whom 73% authored $\leq$ 10 tweets each. We extracted the two following subsets:

1. Prol – all tweets $(28,625,352)$ of all users $(8,262)$ who authored $\geq$ $2,000$ tweets. We refer to them as highly-prolific tweeters.
2. Low – all tweets $(3,449,635)$ authored by $10,000$ users, randomly chosen among all users $(73,004)$ who authored $[300, 400]$ tweets. Since the original dataset is from 2009, given ever-increasing popularity of Twitter, we speculate that Low is a sample of very large demographic.

We extracted two subsets in order to assess linkability of tweets for users at different prolificacy scales. Note that retweets have not been filtered out, as they are considered part of users contributions. In addition, URLs and user-mentions were removed from tweets.

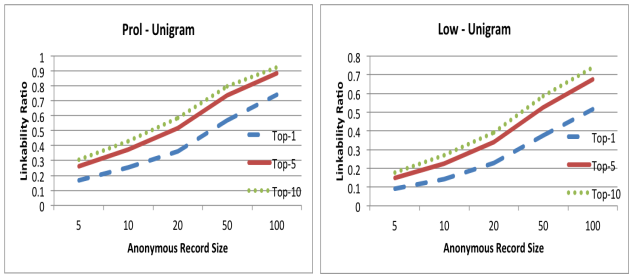
Setting: We adopt linkability-related definitions and abbreviations from [2]. For each author U in both Prol and Low, we randomly split U's tweets into two sets: Identified Record (IR), and Anonymous Record (AR). The set of all IRs is used for training our matching model and the set of all ARs is used to assess linkability – accuracy of linking an AR to its corresponding IR. For each U, we use the matching model to link U's AR to a corresponding IR by returning a list of candidate IRs, sorted in decreasing order of likelihood of being the correct match.

A given AR is considered to have Top-x linkability if the actual corresponding IR is among the top x candidate IR records that the model returns. Performance is measured in terms of linkability ratio (LR), the percentage of ARs that have been correctly classified within the top x candidates. Three x values were considered in the experiments: 1, 5, and 10.

Each U's tweets were partitioned into IR and AR as follows: First, all U's tweets are randomly shuffled. Then, all, except the last 100, tweets are assigned to IR. For the last 100 tweets, first y are assigned to AR. The value of y varies among: 5, 10, 20, 50 and 100.

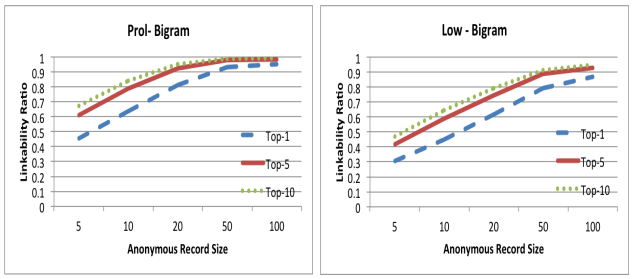
[1]We estimate all probabilities of the form $P(token_i|U)$ using the Maximum Likelihood estimator [4] with Laplace smoothing [8].

[2]Excluding October due to some technical difficulties in extracting data.



(a) (b)

Figure 1: Top-1, Top-5, and Top-10 LRs of unigram NB model for Prol (a) and Low (b)



(a) (b)

Figure 2: Top-1, Top-5, and Top-10 LRs of bigram NB model for Prol (a) and Low (b)

4. LINKABILITY ANALYSIS

We use the NB model to link tweets based on two simple lexical token types: (1) unigrams: all letters in the English alphabet – 26 total, and (2) bigrams: all possible two-letter combinations – 676 total. We perform separate analyses on Prol and Low, and compare respective results.

4.1 Unigrams and Bigrams

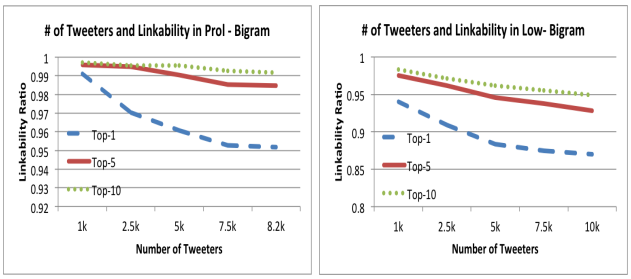
Unigrams. Figures 1(a) and 1(b) show Top-1, Top-5 and Top-10 LRs for variable AR sizes in Prol and Low, respectively. First, we observe that all curves exhibit a clear upward trend, showing that the bigger the AR, the higher LR is obtained. For AR of 100, Top-10 LR is as high as 92% for Prol and 74% for Low. Relatively high LRs are observed for Top-1(Top-5) – 74% (88%) of ARs are linked in Prol for AR of 100.

Bigrams. Similarly, Figures 2(a) and 2(b) show Top-1, Top-5 and Top-10 LRs, for the case of bigrams. Substituting unigrams with bigrams substantially increases LRs, even with a rather small AR. For instance, Top-1 LR is 95% and 87% for Prol and Low, respectively, for AR of 100. Even for small AR sizes, high LRs are achieved; e.g., for AR size of 5(10), Top-10 LR exceeds 67(84)% in Prol.

4.2 Varying the Number of Users

Thus far, we considered the full set of users: $8,262$ in Prol and $10,000$ in Low. We now reduce the number of users in both sets and explore effect on linkability. We use bigrams

since they outperform unigrams. The number of users in Prol and Low is varied between 1,000 and full set size, i.e., 8,262 and 10,000, respectively. Figures 3(a) and 3(b) show LRs for varying set sizes with AR of 100. Interestingly, LR does not decrease much as the set size increases. For example, looking at the entire range, Top-1 LR does not decrease over 4% and 7% in Prol and Low, respectively. This shows that the bigram-based NB model is very resilient against increasing the number of users. We believe that this should be worrisome to tweeters who use multiple accounts.



(a) (b)

Figure 3: LRs of bigram NB model, varying # users in Prol (a) and Low (b)

4.3 Improving Unigram Model

We already established that relying only on bigrams yields very high LRs. However, bigrams require more resources than unigrams: 676 vs 26 tokens. Thus, bigram-based models are less scalable. To this end, we consider improving LRs using only unigrams, by combining them with the unigrams of hashtags. Hashtags are peculiar, yet popular, feature of Twitter. Not surprisingly, many tweets in our dataset contain one or more hashtags. We first filter out from Prol and Low all tweets that do not include any hashtags. We then discard all users with fewer than 300 hashtag-containing tweets; this is in order to populate their corresponding AR sets with 100 tweets. This leaves us with 3,179 and 160 users in Prol and Low, respectively. Since the resultant size of Low (in terms of users) is small, we confine our analysis to Prol. The number of tweets in filtered Prol is 4,274,188. We use unigram tokens within hashtags, which include 11 non-alphabetical characters (i.e., $0-9$ and "_"), thus ending up with 37 tokens.

We then try to improve LRs by combining unigrams of tweets without hashtags (26 tokens) with hashtags-based unigrams (37 tokens). This yields a total of 63 tokens, which is still much less than 676 with bigrams.

As discussed in Section 2, we use $\sum_i logP(t_i|U)$ $Log-Sum$ to sort the matching users from a set of tokens. Now, we use a weighted average to combine the $Log-Sum$ of tweet-text-based and hashtag-based tokens as follows:

$$(\beta)\times \sum_{t_i^{tweet}} logP(t_i^{tweet}|U)+(1-\beta)\times \sum_{t_i^{hashtag}} logP(t_i^{hashtag}|U)$$

However, there is no clear way to assign a value to β. We experimented with several choices. Specifically, we tried all β values ranging in $[0-1]$ in 0.1 increments and observed the highest LR with $\beta = 0.6$ in Top-1, Top-5 and Top-10 LRs. Thus, 0.6 was selected for the combined model[3]. Figure 4 shows Top-1 LR for combining unigrams of tweet-texts and hashtags for the filtered Prol. This combination boosts LRs by $8-13\%$.

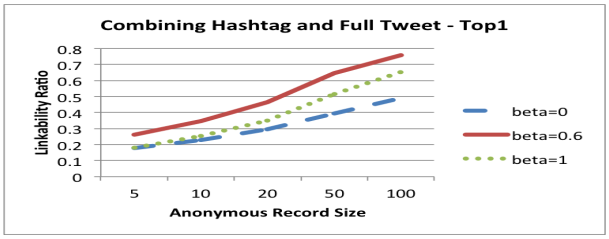


Figure 4: Top-1, Top-5 and Top-10 LRs with different β values

4.4 Dual-Account Tweeters

So far, we analyzed sets of tweets, each authored by a distinct user corresponding to a Twitter account. After *artificially* splitting each such set into Identification Record (IR) and Anonymous Record (AR), we discovered that they are highly linkable. In practice, our technique aims to link **distinct users**, i.e., multiple bodies of tweets emanating from different Twitter accounts.

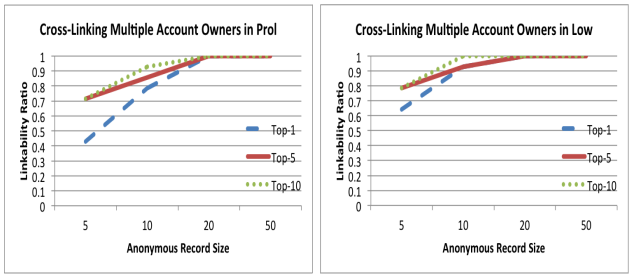
Clearly, we can not assemble a comprehensive collection of all multi-account sets of tweets, even for a fixed period of time. Instead, as ground truth, we use a fairly small number of account-pairs that are known to be operated by the same author/user. In order to collect tweets from dual-account authors, we manually and extensively searched the web for public information pertaining to individuals who operate multiple accounts. For example, we used search variations of keywords, such as "multiple twitter accounts", and found several blog posts where Twitter users publicly reveal tweeting via multiple accounts and list these accounts. We assembled a set of 28 accounts belonging to 14 distinct users, each with 2 accounts. We refer to the set of tweets from these accounts as Dual dataset.

We again use bigram NB as the linking model for the Dual dataset. As a sanity check, we first assessed NB's performance with Dual without mixing in tweets from any external datasets. For each dual-account user, we randomly selected one of the two accounts (i.e, all tweets therein) as IR, and the other – as AR. AR size was varied between 5 and 50, by randomly selecting tweets[4]. Top-1 LR is 100% for AR $\geq$ 20. Also, Top-1 LR exceeds 85% for AR < 10 [5].

[3]Note that β selection process (training) is restricted to the set of IRs (ARs are excluded). That is, β is learned using IRs, by further splitting them into identified and anonymized sets, and computing the corresponding LR values in the filtered version of Prol.

[4]Maximum AR size was 50, instead of earlier 100, since some accounts had < 100 tweets.

[5]Due to space limitations, we omit the corresponding figure; see [7].



(a) (b)

Figure 5: Top-1, Top-5 and Top-10 LRs in Dual when merging accounts with Prol – (a) and Low – (b)

As the next step, we verify that the model is scalable for dual-account owners, i.e., performs well if tweets from Dual are merged with Prol and Low datasets, respectively. We merge IRs from Dual with Prol / Low. Likewise, we augment ARs of Dual with ARs of Prol / Low. Figures 5(a) and 5(b) show LRs of 14 dual-account owners in Dual augmented by Prol and Low, respectively. Once again, Top-1, Top-5 and Top-10 are at 100% when AR size exceeds 20. This clearly confirms efficacy of the NB bigram model.

5. RELATED WORK

Author Attribution in Twitter: Some prior results focused on authorship identification and stylometric analysis of microblogging. [10] considered re-identifying authorship of tweets from a set of three authors, while using over $5,000$ dimensions as input for a Support Vector Machine (SVM) classifier. Similarly, [6] investigated pseudonymity for a set of 50 Twitter users. [5] studied the use of n-grams with NB as the linkability model. In particular, 2- to 6-grams are evaluated and a 98% linkability is achieved in the setting of 50 authors. An identification technique based on extracting a set of lexical and syntactical features along with SVM is proposed in [3]. It achieves 91% accuracy for a set of 15 authors. There are four main differences between our results work and aforementioned studies. First, we assess linkability on a large scale, i.e., the number of tweeters is much larger than in prior work. Second, we use unigrams, which drastically reduces the number of tokens. Third, we include hashtags and show their efficacy in linkability when used alone or in combination with other unigrams. Finally, we successfully re-produce and confirm linkability results for a small set of actual dual-account tweeters.

General Author Attribution. One of the best-known results is [1], which uses an extensive set of features, called Writeprints, and a technique that is based on Karhunen-Loeve-transforms. This approach attains identification accuracy of 91%. Other author attribution studies for user-review sites and blogs are [2] and [9]. A survey of this topic can be found in [11]. Also, a more extensive overview of related work appears in the full version of this paper [7].

6. SUMMARY & ACKNOWLEDGMENTS

This paper reports on a large-scale linkability analysis of tweets, based on two datasets, each consisting of $> 8,000$ tweeters. As shown, tweets are highly linkable, even using simple unigram and/or bigram distributions. This is confirmed by linkability analysis of a set of tweets that includes actual dual-account users. These are worrisome results for privacy-conscious tweeters.

G. Tsudik and E. Oguz were supported by NSF CSR Award 1213140. M. Almishari's research was supported in part by King Saud University. NICTA is funded by the Australian Government as represented by the Department of Broad-band, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

7. REFERENCES

[1] A. Abbasi and H. Chen. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. In *ACM Transactions on Information Systems*, 2008.

[2] M. Almishari and G. Tsudik. Exploring linkability in user reviews. In *European Symposium on Research in Computer Security*, 2012.

[3] M. Bhargava, P. Mehndiratta, and K. Asawa. Stylometric Analysis for Authorship Attribution on Twitter. In *Big Data Analytics*, 2013.

[4] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] S. R. Boutwell. Authorship Attribution of Short Messages Using Multimodal Features. In *Master Thesis, Naval Postgraduate School*, 2011.

[6] R. Layton, P. Watters, and R. Dazeley. Authorship Attribution for Twitter in 140 Characters or Less. In *Cybercrime and Trustworthy Computing Workshop (CTC)*, 2010.

[7] M. A. Mishari, D. Kaafar, G. Tsudik, and E. Oguz. Are 140 characters enough? A large-scale linkability study of tweets. *CoRR*, abs/1406.2746, 2014.

[8] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[9] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. Shin, and D. Song. On the Feasibility of Internet-Scale Author Identification. In *IEEE Symposium on Security and Privacy*, 2012.

[10] R. Silva and G. Laboreiro and L. Sarmento and T. Grant and E. Oliveira and B. Maia. Automatic Authorship Analysis of Micro-Blogging Messages. In *NLDB*, 2011.

[11] E. Stamatatos. A Survey of Modern Authorship Attribution Methods. In *Journal of the American Society for Information Science and Technology*, 2009.

[12] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, 2011.