Authorship Attribution of Micro-Messages

Roy Schwartz, Oren Tsur, Ari Rappoport
Institute of Computer Science
Hebrew University of Jerusalem
{roys02|oren|arir}@cs.huji.ac.il

Moshe Koppel
Department of Computer Science
Bar Ilan University
koppel@macs.biu.ac.il

Abstract

Work on authorship attribution has traditionally focused on long texts. In this work, we tackle the question of whether the author of a very short text can be successfully identified. We use Twitter as an experimental testbed. We introduce the concept of an author's unique "signature", and show that such signatures are typical of many authors when writing very short texts. We also present a new authorship attribution feature ("flexible patterns") and demonstrate a significant improvement over our baselines. Our results show that the author of a single tweet can be identified with good accuracy in an array of flavors of the authorship attribution task.

1 Introduction

Research in authorship attribution has developed substantially over the last decade (Stamatatos, 2009). The vast majority of such research has been dedicated towards finding the author of long texts, ranging from single passages to book chapters. In recent years, the growing popularity of social media has created special interest, both theoretical and computational, in short texts. This has led to many recent authorship attribution projects that experimented with web data such as emails (Abbasi and Chen, 2008), web forum messages (Solorio et al., 2011) and blogs (Koppel et al., 2011b). This paper addresses the question to what extent the authors of very short texts can be identified. To answer this question, we experiment with Twitter tweets.

Twitter messages (tweets) are limited to 140 characters. This restriction imposes major difficulties on authorship attribution systems, since authorship attribution methods that work well on long texts are often not as useful when applied to short texts (Burrows, 2002; Sanderson and Guenter, 2006).

Nonetheless, tweets are relatively self-contained and have smaller sentence length variance compared to excerpts from longer texts (see Section 3). These characteristics make Twitter data appealing as a testbed when focusing on short texts. Moreover, an authorship attribution system of tweets may have various applications. Specifically, a range of cybercrimes can be addressed using such a system, including identity fraud and phishing.

In this paper, we introduce the concept of k-*signatures*. We denote the k-*signatures* of an author a as the features that appear in at least $k\%$ of a's training samples, while not appearing in the training set of any other author. When k is large, such signatures capture a unique style used by a. An analysis of our training set reveals that unique k-signatures are typical of many authors. Moreover, a substantial portion of the tweets in our training set contain at least one such signature. These findings suggest that a single tweet, although short and sparse, often contains sufficient information for identifying its author. Our results show that this is indeed the case.

We train an SVM classifier with a set of features that include character n-grams and word n-grams. We use a rigorous experimental setup, with varying number of authors (values between 50-1,000) and various sizes of the training set, ranging from 50 to 1,000 tweets per author. In all our experiments, a single tweet is used as test document. We also use a setting in which the system is allowed to respond *don't know* in cases of uncertainty. Applying this option results in higher precision, at the expense of

lower recall.

Our results show that the author of a tweet can be successfully identified. For example, when using a dataset of as many as 1,000 authors with 200 training tweets per author, we are able to obtain 30.3% accuracy (as opposed to a random baseline of only 0.1%). Using a dataset of 50 authors with as few as 50 training tweets per author, we obtain 50.7% accuracy. Using a dataset of 50 authors with 1,000 training tweets per author, our results reach as high as 71.2% in the standard classification setting, and exceed 91% accuracy with 60% recall in the *don't know* setting.

We also apply a new set of features, never previously used for this task – flexible patterns. Flexible patterns essentially capture the context in which function words are used. The effectiveness of function words as authorship attribution features (Koppel et al., 2009) suggests using flexible pattern features. The fact that flexible patterns are learned from plain text in a fully unsupervised manner makes them domain and language independent. We demonstrate that using flexible patterns gives significant improvement over our baseline system. Furthermore, using flexible patterns, our system obtains a 6.1% improvement over current state-of-the-art results in authorship attribution on Twitter.

To summarize, the contribution of this paper is threefold.

- We provide the most extensive research to date on authorship attribution of micro-messages, and show that authors of very short texts can be successfully identified.
- We introduce the concept of an author's unique k-signature, and demonstrate that such signatures are used by many authors in their writing of micro-messages.
- We present a new feature for authorship attribution – flexible patterns – and show its significant added value over other methods. Using this feature, our system obtains a 6.1% improvement over the current state-of-the-art.

The rest of the paper is organized as follows. Sections 2 and 3 describe our methods and our experimental testbed (Twitter). Section 4 presents the concept of k-signatures. Sections 5 and 6 present our experiments and results. Flexible patterns are presented in Section 7 and related work is presented in Section 8.

2 Methodology

In the following we briefly describe the main features employed by our system. The features below are binary features.

Character n-grams. Character n-gram features are especially useful for authorship attribution on micro-messages since they are relatively tolerant to typos and non-standard use of punctuation (Stamatatos, 2009). These are common in the non-formal style generally applied in social media services. Consider the example of misspelling "Britney" as "Brit**t**ney". The misspelled name shares the 4-grams "Brit" and "tney" with the correct name. As a result, these features provide information about the author's style (or at least her topic of interest), which is not available through lexical features.

Following standard practice, we use 4-grams (Sanderson and Guenter, 2006; Layton et al., 2010; Koppel et al., 2011b). White spaces are considered characters (i.e., a character n-gram may be composed of letters from two different words). A single white-space is appended to the beginning and the end of each tweet. For efficiency, we consider only character n-gram features that appear at least t_{cng} times in the training set of at least one author (see Section 5).

Word n-grams. We hypothesize that word n-gram features would be useful for authorship attribution on micro-messages. We assume that under a strict length restriction, many authors would prefer using short, repeating phrases (word n-grams).

In our experiments, we consider $2 \leq n \leq 5$.[1] We regard sequences of punctuation marks as words. Two special words are added to each tweet to indicate the beginning and the end of the tweet. For efficiency, we consider only word n-gram features that appear at least t_{wng} times in the training set of at least one author (see Section 5).

Model. We use libsvm's Matlab implementation of a multi-class SVM classifier with a linear kernel

[1] We skip unigrams as they are generally captured by the character n-gram features.

(Chang and Lin, 2011). We use ten-fold cross validation on the training set to select the best regularization factor between 0.5 and 0.005.[2]

3 Experimental Testbed

Our main research question in this paper is to determine the extent to which authors of very short texts can be identified. A major issue in working with short texts is selecting the right dataset. One approach is breaking longer texts into shorter chunks (Sanderson and Guenter, 2006). We take a different approach and experiment with micro-messages (specifically, tweets).

Tweets have several properties making them an ideal testbed for authorship attribution of short texts. First, tweets are posted as single units and do not necessarily refer to each other. As a result, they tend to be self contained. Second, tweets have more standardized length distribution compared to other types of web data. We compared the mean and standard deviation of sentence length in our Twitter dataset and in a corpus of English web data (Ferraresi et al., 2008).[3] We found that (a) tweets are shorter than standard web data (14.2 words compared to 20.9), and (b) the standard deviation of the length of tweets is much smaller (6.4 vs. 21.4).

Pre-Processing. We use a Twitter corpus that includes approximately 5×10^8 tweets.[4] All non-English tweets and tweets that contain fewer than 3 words are removed from the dataset. We also remove tweets marked as retweets (using the RT sign, a standard Twitter symbol to indicate that this tweet was written by a different user). As some users retweet without using the RT sign, we also remove tweets that are an exact copy of an existing tweet posted in the previous seven days.

Apart from plain text, some tweets contain references to other Twitter users (in the format of @<user>). Since using reference information makes this task substantially easier (Layton et al., 2010), we replace each user reference with the special meta tag REF. For sparsity reasons, we also replace web addresses with the meta tag URL, numbers with the meta tag NUM, time of day with the meta tag TIME and dates with the meta tag DATE.

[2] In practice, 0.05 or 0.1 are selected in almost all cases.

[3] http://wacky.sslmit.unibo.it

[4] These comprise ∼15% of all public tweets created from May 2009 to March 2010.

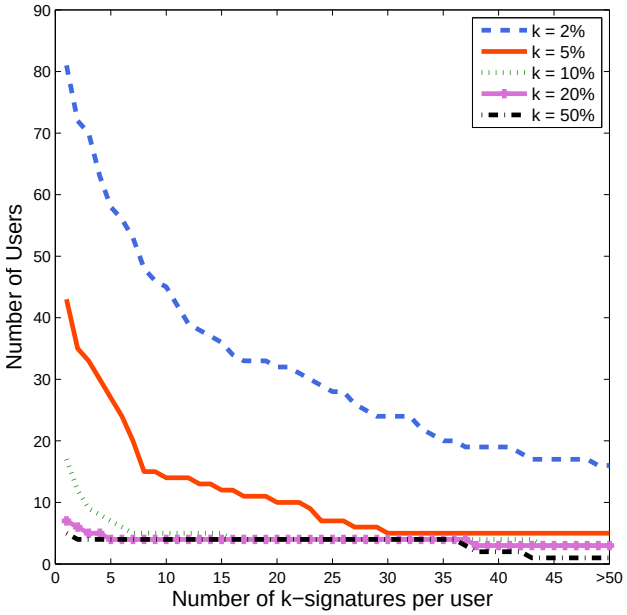

Figure 1: Number of users with at least x k-signatures (100 authors, 180 training tweets per author).

4 k-Signatures

In this section, we show that many authors adopt a unique style when writing micro-messages. This style can be detected by a strong classification algorithm (such as SVM), and be sufficient to correctly identify the author of a single tweet.

We define the concept of the *k-signature* of an author a to be a feature that appears in at least $k\%$ of a's training set, while not appearing in the training set of any other user. Such signatures can be useful for identifying future (unlabeled) tweets written by a.

To validate our hypothesis, we use a dataset of 100 authors with 180 tweets per author. We compute the number of k-signatures used by each of the authors in our dataset. Figure 1 shows our results for a range of k values (2%, 5%, 10%, 20% and 50%). Results demonstrate that 81 users use at least one 2%-signature, 43 users use at least one 5%-signature, and 17 users use at least one 10%-signature. These results indicate that a large portion of the users adopt a unique signature (or set of signatures) when writing short texts. Table 1 provides examples of 10%-signatures.

Signature Type	10%-signature	Examples
Character n-grams	‘ ^_^’	REF oh ok ^_^ Glad you found it!
		Hope everyone is having a good afternoon ^_^
		REF Smirnoff lol keeping the goose in the freezer ^_^
	‘**yew** ’	gurl **yew** serving me tea nooch
		REF about wen **yew** and ronnie see each other
		REF lol so **yew** goin to check out tini’s tonight huh???
Word n-grams	**.. lal**	REF aww those are cool where u get those.. how do ppl react**.. lal**
		Ludas album is gone be hott**.. lal**
		Dayum refs don’t get injury timeouts**.. lal**.. get him off the field..
	smoochies , e3	I’m just back after takin’ a very long, icy cold shower.....Shivering **smoochies,E3** http://bit.ly/4CzzP9
		A blue stout or two would be nice as well, Purr!Blue smooth **smoochies,E3** http://bit.ly/75D4fO
		That is sooooooooooooooooooo unfair!Double **smoochies,E3** http://bit.ly/07sXRGX

Table 1: Examples of 10%-signatures.

Results also show that seven users use one or more 20%-signatures, and five users even use one or more 50%-signatures. Looking carefully at these users, we find that they write very structured messages, and are probably bots, such as news feeds, bidding systems, etc. Table 2 provides examples of tweets posted by such users.[5]

Another interesting question is how many tweets contain at least one k-signature. Figure 2 shows for each user the number of tweets in her training set for which at least one k-signature is found. Results demonstrate that a total of 18.6% of the training tweets contain at least one 2%-signature, 10.3% the training tweets contain at least one 5%-signature and 6.5% of the training tweets contain at least one 10%-signature. These findings validate our assumption that many users use k-signatures in short texts.

These findings also have direct implications on authorship attribution of micro-messages, since k-signatures are reliable classification features. As a result, texts written by authors that tend to use k-signatures are likely to be easily identified by a reasonable classification algorithm. Consequently, k-signatures provide a possible explanation for the high quality results presented in this paper.

In the broader context, the presence (and contribution) of k-signatures is in line with the hypothesis proposed by (Davidov et al., 2010a): while still using an informal and unstructured (grammatical) language, authors tend to use typical and unique structures in order to allow a short message to stand alone without a clear conversational context.

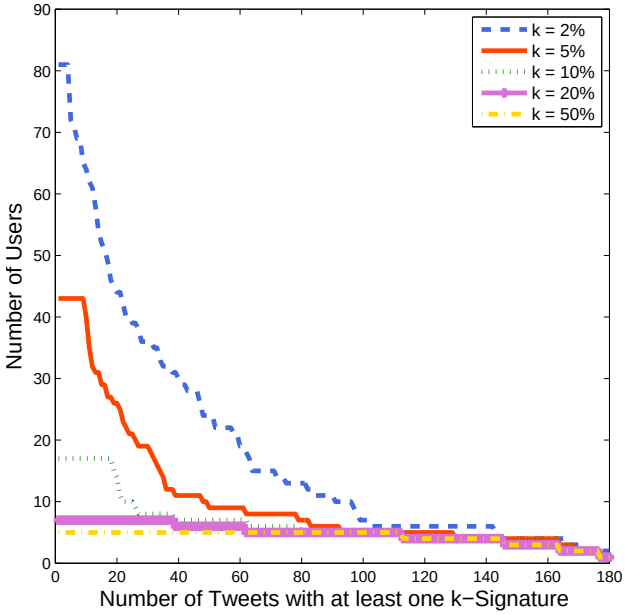

Figure 2: Number of users with at least x training tweets that contain at least one k-signature (100 authors, 180 training tweets per author).

[5]Our k-signature method can actually be useful for automatically identifying such users. We defer this to future work.

User	20%-signature	Examples
1	**I'm listening to :**	**I'm listening to:** Sigur R?s ? Intro: http://www.last.fm/music/Sigur+R%C3%B3s http://bit.ly/3XJHyb
		I'm listening to: Tina Arena ? In Command: http://www.last.fm/music/Tina+Arena http://bit.ly/7q9E25
		I'm listening to: Midnight Oil ? Under the Overpass: http://www.last.fm/music/Midnight+Oil http://bit.ly/7IH4cg
2	**news now (str)**	#Hotel **News Now(STR)** 5 things to know: 27 May 2009: From the desks of the HotelNewsNow.com editor... http://bit.ly/aZTZOq #Tourism #Lodging
		#Hotel **News Now(STR)** Five sales renegotiating tactics: As bookings representatives press to reneg... http://bit.ly/bHPn2L
		#Hotel **News Now(STR)** Risk of hotel recession retreats: The Hotel Industry's Pulse Index increases... http://bit.ly/a8EKrm #Tourism #Lodging
3	**(NUM bids) end date :**	NEW PINK NINTENDO DS LITE CONSOLE WITH 21 GIFTS + CASE: £66.50 **(13 Bids) End Date:** Tuesday Dec-08-2009 17:.. http://bit.ly/7uPt6V
		Microsoft Xbox 360 Game System - Console Only - Working: US $51.99 **(25 Bids) End Date:** Saturday Dec-12-2009 13:.. http://bit.ly/8VgdTv
		Microsoft Sony Playstation 3 (80 GB) Console 6 Months Old: £190.00 **(25 Bids) End Date:** Sunday Dec-13-2009 21:21:39 G.. http://bit.ly/7kwtDS

Table 2: Examples of tweets published by very structured users, suspected to be bots, along with one of their 20%-signatures.

5 Experiments

We report of three different experimental configurations. In the experiments described below, each dataset is divided into training and test sets using ten-fold cross validation. On the test phase, each document contains a single tweet.

Experimenting with varying Training Set Sizes. In order to test the affect of the training set size, we experiment with an increasingly larger number of tweets per author. Experimenting with a range of training set sizes serves two purposes: (a) to check whether the author of a tweet can be identified using a very small number of (short) training samples, and (b) check how much our system can benefit from training on a larger corpus.

In our experiments we only consider users who posted between 1,000–2,000 tweets[6] (a total of 10,183 users), and randomly select 1,000 tweets per user. From these users, we select 10 groups of 50 users each.[7] We perform a set of classification experiments, selecting for each author an increasingly larger subset of her 1,000 tweets as training set. Subset sizes are (50, 100, 200, 500, 1,000). Threshold values for our features in each setting (see Section 2) are $(2, 2, 4, 10, 20)$ for t_{cng} and $(2, 2, 2, 3, 5)$ for t_{wng}, respectively.

Experimenting with varying Numbers of Authors. In a second set of experiments, we use an increasingly larger number of authors (values between 100-1,000), in order to check whether the author of a very short text can be identified in a "needle in a haystack" type of setting.

Due to complexity issues, we only experiment with 200 tweets per author as training set. We select groups of size 100, 200, 500 and 1,000 users (one group per size). We use the same threshold values as the 200 tweets per author setting previously described (t_{cng} = 4, t_{wng} = 2).

[6] This range is selected since on one hand we want at least 1,000 tweets per author for our experiments, and on the other hand we noticed that users with a larger number of tweets in corpus tend to be spammers or bots that are very easy to identify, so we limit this number to 2,000.

[7] An eleventh group is selected as development set.

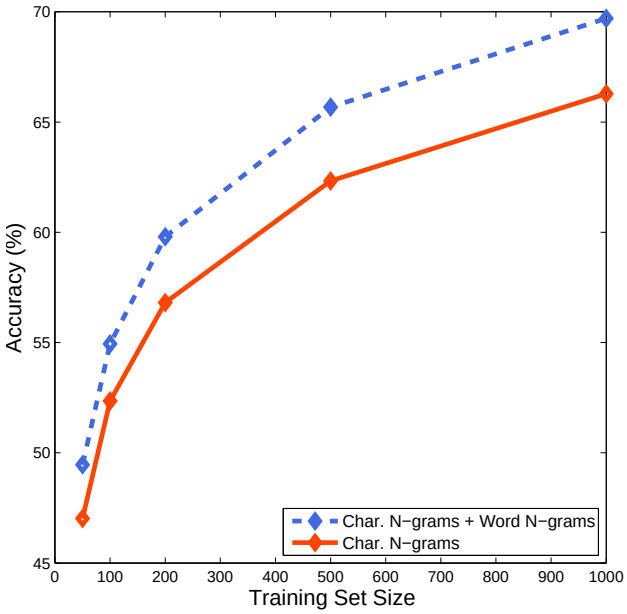

Figure 3: Authorship attribution accuracy for 50 authors with various training set sizes. The values are averaged over 10 groups. The random baseline is 2%.

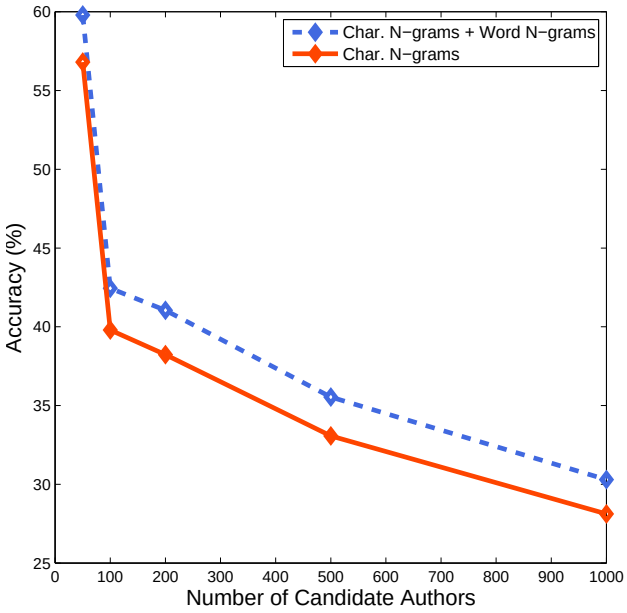

Figure 4: Authorship attribution accuracy with varying number of candidate authors, using 200 training tweets per author. The random baselines for 50[9], 100, 200, 500 and 1,000 authors are 2%, 1%, 0.5%, 0.2% and 0.1%, respectively.

Recall-Precision Tradeoff. Another aspect of our research question is the level of certainty our system has when suggesting an author for a given tweet. In cases of uncertainty, many real life applications would prefer not to get any response instead of getting a response with low certainty. Moreover, in real life applications we are often not even sure that the real author is part of our training set. Consequently, we allow our system to respond "*don't know*" in cases of low confidence (Koppel et al., 2006; Koppel et al., 2011b). This allows our system to obtain higher precision, at the expense of lower recall.

To implement this feature, we use SVM's probability estimates, as implemented in libsvm. These estimates give a score to each potential author. These scores reflect the probability that this author is the correct author, as decided by the prediction model. The selected author is always the one with the highest probability estimate.

As selection criterion, we use a set of increasingly larger thresholds (0.05-0.9) for the probability of the selected author. This means that we do not select test samples for which the selected author has a probability estimate value lower than the threshold.

6 Basic Results

Experimenting with varying Training Set Sizes. Figure 3 shows results for our experiments with 50 authors and various training set sizes. Results demonstrate that authors of very short texts can be successfully identified, even with as few as 50 tweets per author (49.5%). When given more training samples, authors are identified much more accurately (up to 69.7%). Results also show that, according to our hypothesis, word n-gram features substantially improve over character n-grams features only (3% averaged improvement over all settings).

Experimenting with varying Numbers of Authors. Figure 4 shows our results for various numbers of authors, using 200 tweets per author as training set. Results demonstrate that authors of an unknown tweet can be identified to a large extent even when there are as many as 1,000 candidate authors (30.3%, as opposed to a random baseline of only 0.1%). Results further validate that word n-gram features substantially improve over character

[9]Results for 50 authors with 200 tweets per author are taken from Figure 3.

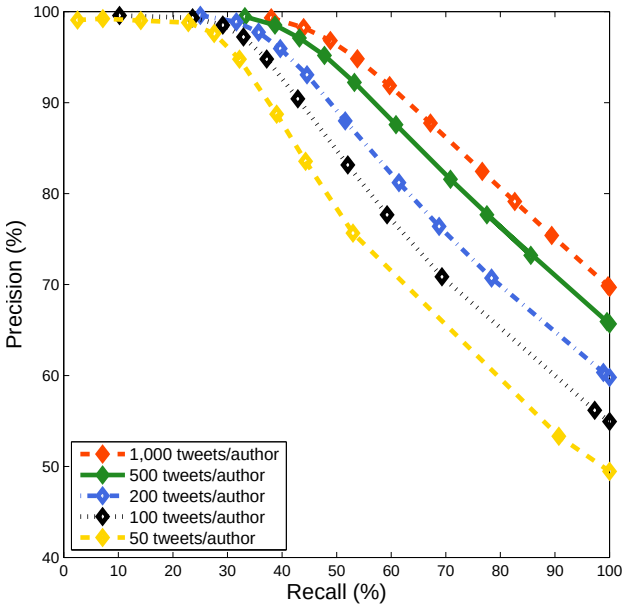

Figure 5: Recall-precision curves for 50 authors with varying training set sizes.

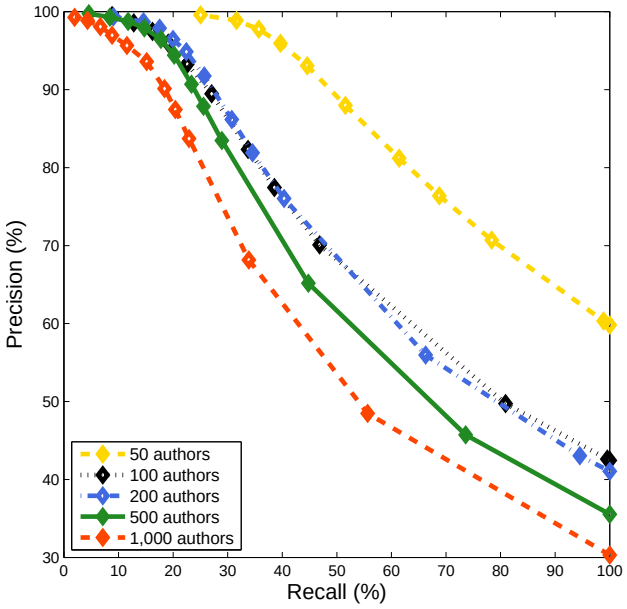

Figure 6: Recall-precision curves for varying number of authors.

n-grams features (2.6% averaged improvement).

Recall-Precision Tradeoff. Figure 5 shows the recall-precision curves for our experiments with 50 authors and varying training set sizes. Results demonstrate that we are able to obtain very high precision (over 90%) while still maintaining a relatively high recall (from ∼35% recall for 50 tweets per author up to $> 60\%$ recall for 1,000 tweets per author).

Figure 6 shows the recall-precision curves for our experiments with varying number of authors. Results demonstrate that even in the 1,000 authors setting, we are able to obtain high precision values (90% and 70%) with reasonable recall values (18% and ∼30%, respectively).

7 Flexible Patterns

In previous sections we provided strong evidence that authors of micro-messages can be successfully identified using standard methods. In this section we present a new feature, never previously used for this task – flexible patterns. We show that flexible patterns can be used to improve classification results.

Flexible patterns are a generalization of word n-grams, in the sense that they capture potentially unseen word n-grams. As a result, flexible patterns can pick up fine-grained differences between authors' styles. Unlike other types of pattern features, flexible patterns are computed automatically from plain text. As such, they can be applied to various tasks, independently of domain and language. We describe them in detail.

Word Frequency. Flexible patterns are composed of high frequency words (HFW) and content words (CW). Every word in the corpus is defined as either HFW or CW. This clustering is performed by counting the number of times each word appears in the corpus of size s. A word that appears more than $10^{-4} \times s$ times in a corpus is considered HFW. A word that appears less than $10^{-3} \times s$ times in a corpus is considered CW. Some words may serve both as HFWs and CWs (see Davidov and Rappoport (2008b) for discussion).

Structure of a Flexible Pattern. Flexible patterns start and end with an HFW. A sequence of zero or more CWs separates consecutive HFWs. At least one CW must appear in every pattern.[10] For efficiency, at most six HFWs (and as a result, five CW sequences) may appear in a flexible pattern. Examples of flexible patterns include

1. "the_{HFW} CW of_{HFW} the_{HFW}"

[10] Omitting this treats word n-grams as flexible patterns.

Flexible Pattern Features. Flexible patterns can serve as binary classification features; a tweet matches a given flexible pattern if it contains the flexible pattern sequence. For example, (1) is matched by (2).

2. "*Go to* $\underline{the_{HFW}\ house_{CW}\ of_{HFW}\ the_{HFW}}$ *rising sun*"

Partial Flexible Patterns. A flexible pattern may appear in a given tweet with additional words not originally found in the flexible pattern, and/or with only a subset of the HFWs (Davidov et al., 2010a). For example, (3) is a partial match of (1), since the word "great" is not part of the original flexible pattern. Similarly, (4) is another partial match of (1), since (a) the word "good" is not part of the original flexible pattern and (b) the second occurrence of the word "the" does not appear in (4) (missing word is marked by ___).

3. "$\underline{The_{HFW}\ \mathbf{great}_{\boldsymbol{HFW}}\ king_{CW}\ of_{HFW}\ the_{HFW}}$ *ring*"

4. "$\underline{The_{HFW}\ \mathbf{good}_{\boldsymbol{HFW}}\ king_{CW}\ of_{HFW}\ \_\_}$ *Spain*"

We use such cases as features with lower weight, proportional to the number of found HFWs in the tweet ($w = \frac{0.5 \times n_{found}}{n_{expected}}$). For example, (1) receives a weight of 1 (complete match) against (2). Against (3), it receives a weight of 0.5 ($= \frac{0.5 \times 3}{3}$, partial match with no missing HFWs). Against (4) it receives a weight of $1/3$ ($= \frac{0.5 \times 2}{3}$, partial match with only 2/3 HFWs found).

Experimenting with Flexible Pattern Features. We repeat our experiments with varying training set sizes (see Section 5) with two more systems: one that uses character n-grams and flexible pattern features, and another that uses character n-grams, word n-grams and flexible patterns. High frequency word counts are computed separately for each author using her training set. We only consider flexible pattern features that appear at least t_{fp} times in the training set of at least one author. Values of t_{fp} for training set sizes (50, 100, 200, 500, 1,000) are (2, 3, 7, 7, 8), respectively.

Results. Figure 7 shows our results. Results demonstrate that flexible pattern features have an added value over both character n-grams alone (averaged 2.9% improvement) and over character n-grams and word n-grams together (averaged 1.5% improvement). We perform t-tests on each of our training set sizes to check whether the latter improvement is significant. Results demonstrate that it is highly significant in all settings, with p-values smaller than values between 10^{-3} (for 50 tweets per author) and 10^{-8} (1,000 tweets per author).

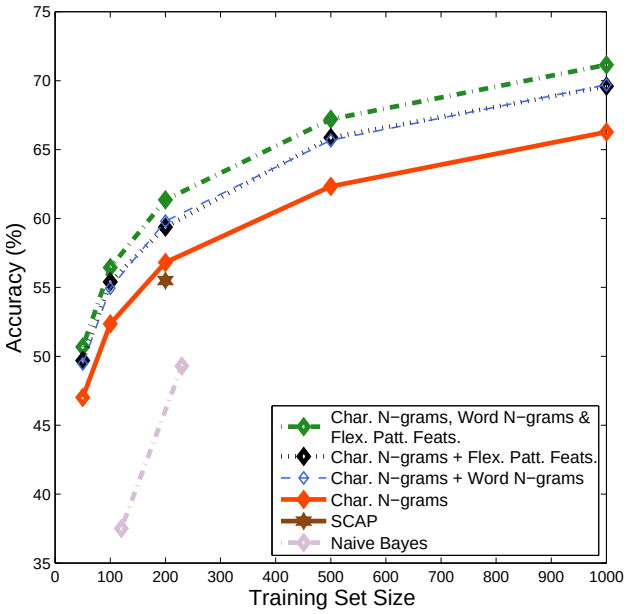

Figure 7: Authorship attribution accuracy for 50 authors with various training set sizes and various feature sets. The values are averaged over 10 groups. The random baseline is 2%.
Comparison to previous work: SCAP – SCAP algorithm results, as reported by (Layton et al., 2010), Naive Bayes – Naive Bayes algorithm results, as reported by (Boutwell, 2011).

Comparison to Previous Works. Figure 7 also shows results for the only two works that experimented in some of the settings we experimented in: Layton et al. (2010) and Boutwell (2011) (see Section 8). Our system substantially outperforms these two systems, by margins of 5.9% to 19%. These margins are explained by the choice of algorithm (SVM and not SCAP/naive Bayes) and our set of features (character n-grams + word n-grams + flexible patterns compared to character n-grams only). In order to rule out the possibility that these margins stem from using different datasets, we tested our system on the dataset used in (Layton et al., 2010). Our system obtains even higher results on this dataset than on our datasets (61.6%, a total im-

provement of 6.1% over (Layton et al., 2010)).

Discussion. To illustrate the additional contribution of flexible patterns over word n-grams, consider the following tweets, written by the same author.

5. "... $\underline{the_{HFW}\ way_{CW}\ I_{HFW}}$ *treated*$_{CW}$ *her*$_{HFW}$"
6. "... *half of* $\underline{the_{HFW}\ things_{CW}\ I_{HFW}}$ *have seen*"
7. "... $\underline{the_{HFW}\ friends_{CW}\ I_{HFW}}$ *have had for years"*
8. "... *in* $\underline{the_{HFW}\ neighborhood_{CW}\ I_{HFW}}$ *grew up in*"

Consider a case where (5) is part of the test set, while (6-8) appear in the training set. As (5) shares no sequence of words with (6-8), no word n-gram feature is able to identify the author's style in (5). However, this style can be successfully identified using the flexible pattern (9), shared by (5-8).

9. $the_{HFW}\quad CW\quad I_{HFW}$

This demonstrates the added value flexible pattern features have over word n-gram features.

8 Related Work

Authorship attribution dates back to the end of 19th century, when (Mendenhall, 1887) applied sentence length and word length features to plays of Shakespeare. Ever since, many methods have been developed for this task. For recent surveys, see (Koppel et al., 2009; Stamatatos, 2009; Juola, 2012).

Authorship attribution methods can be generally divided into two categories (Stamatatos, 2009). In similarity-based methods, an anonymous text is attributed to some author whose writing style is most similar (by some distance metric). In machine learning methods, which we follow in this paper, anonymous texts are classified, using machine learning algorithms, into different categories (in this case, different authors).

Machine learning papers differ from each other by the features and machine learning algorithm. Examples of features include HFWs (Mosteller and Wallace, 1964; Argamon et al., 2007), character n-gram (Kjell, 1994; Hoorn et al., 1999; Stamatatos, 2008), word n-grams (Peng et al., 2004), part-of-speech n-grams (Koppel and Schler, 2003; Koppel et al., 2005) and vocabulary richness (Abbasi and Chen, 2005).

The various machine learning algorithms used include naive Bayes (Mosteller and Wallace, 1964; Kjell, 1994), neural networks (Matthews and Merriam, 1993; Kjell, 1994), K-nearest neighbors (Kjell et al., 1995; Hoorn et al., 1999) and SVM (De Vel et al., 2001; Diederich et al., 2003; Koppel and Schler, 2003).

Traditionally, authorship attribution systems have mainly been evaluated against long texts such as theater plays (Mendenhall, 1887), essays (Yule, 1939; Mosteller and Wallace, 1964), biblical books (Mealand, 1995; Koppel et al., 2011a) and book chapters (Argamon et al., 2007; Koppel et al., 2007). In recent year, many works focused on web data such as emails (De Vel et al., 2001; Koppel and Schler, 2003; Abbasi and Chen, 2008), web forum messages (Abbasi and Chen, 2005; Solorio et al., 2011), blogs (Koppel et al., 2006; Koppel et al., 2011b) and chat messages (Abbasi and Chen, 2008). Some works focused on SMS messages (Mohan et al., 2010; Ishihara, 2011).

Authorship Attribution on Twitter. The performance of authorship attribution systems on short texts is affected by several factors (Stamatatos, 2009). These factors include the number of candidate authors, the training set size and the size of the test document.

Very few authorship attribution works experimented with Twitter. Unlike our work, all used a single group of authors (group sizes varied between 3-50). Layton et al. (2010) used the SCAP methodology (Frantzeskou et al., 2007) with character n-gram features. They experimented with 50 authors and compared different numbers of tweets per author (values between 20-200). Surprisingly, they showed that their system does not improve when given more training tweets. In our work, we noticed a different trend, and showed that more data can be extremely valuable for authorship attribution systems on micro-messages (see Section 6). Silva et al. (2011) trained an SVM classifier with various features (e.g., punctuation and vocabulary features) on a small dataset of three authors only, with varying training set size. Although their work used a set of Twitter-specific features that we do not explicitly use, our features implicitly cover a large portion of their features (such as punctuation and emoticon

features, which are largely covered by character n-grams).

Boutwell (2011) used a naive Bayes classifier with character n-gram features. She experimented with 50 authors and two training size values (120 and 230). She also provided a set of experiments that studied the effect of joining several tweets into a single document. Mikros and Perifanos (2013) trained an SVM classifier with character n-gram and word n-grams. They experimented with 10 authors of Greek text, and also joined several tweets into a single document. Joining several tweets into a longer document is appealing since it can lead to substantial improvement of the classification results, as demonstrated by the works above. However, this approach requires the test data to contain several tweets that are known a-priori to be written by the same author. This assumption is not always realistic. In our paper, we intentionally focus on a single tweet as document size.

Flexible Patterns. Patterns were introduced by (Hearst, 1992), who used hand crafted patterns to discover hyponyms. Hard coded patterns were used for many tasks, such as discovering meronymy (Berland and Charniak, 1999), noun categories (Widdows and Dorow, 2002), verb relations (Chklovski and Pantel, 2004) and semantic class learning (Kozareva et al., 2008).

Patterns were first extracted in a fully unsupervised manner ("flexible patterns") by (Davidov and Rappoport, 2006), who used flexible patterns in order to establish noun categories, and (Biciçi and Yuret, 2006) who used them for analogy question answering. Ever since, flexible patterns were used as features for various tasks such as extraction of semantic relationships (Davidov et al., 2007; Turney, 2008b; Bollegala et al., 2009), detection of synonyms (Turney, 2008a), disambiguation of nominal compound relations (Davidov and Rappoport, 2008a), sentiment analysis (Davidov et al., 2010b) and detection of sarcasm (Tsur et al., 2010).

9 Conclusion

The main goal of this paper is to measure to what extent authors of micro-messages can be identified. We have shown that authors of very short texts can be successfully identified in an array of authorship attribution settings reported for long documents. This is the first work on micro-messages to address some of these settings. We introduced the concept of *k-signature*. Using this concept, we proposed an interpretation of our results. Last, we presented the first authorship attribution system that uses flexible patterns, and demonstrated that using these features significantly improves over other systems. Our system obtains 6.1% improvement over the current state-of-the-art.

Acknowledgments

We would like to thank Elad Eban and Susan Goodman for their helpful advice, as well as Robert Layton for providing us with his dataset. This research was funded (in part) by the Harry and Sylvia Hoffman leadership and responsibility program (for the first author) and the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20:67–75.

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):7:1–7:29.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 58(6):802–822.

Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proc. of ACL*, pages 57–64, College Park, Maryland, USA.

Ergun Biciçi and Deniz Yuret. 2006. Clustering word pairs to answer analogy questions. In *Proc. of TAINN*, pages 1–8.

Danushka T. Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2009. Measuring the similarity between implicit semantic relations from the web. In *Proc. of WWW*, New York, New York, USA. ACM Press.

Sarah R. Boutwell. 2011. Authorship Attribution of Short Messages Using Multimodal Features. Master's thesis, Naval Postgraduate School.

John Burrows. 2002. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In Dekang Lin and Dekai Wu, editors, *Proc. of EMNLP*, pages 33–40, Barcelona, Spain.

Dmitry Davidov and Ari Rappoport. 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proc. of ACL-Coling*, pages 297–304, Sydney, Australia.

Dmitry Davidov and Ari Rappoport. 2008a. Classification of semantic relationships between nominals using pattern clusters. In *Proceedings of ACL-08: HLT*, pages 227–235, Columbus, Ohio, June. Association for Computational Linguistics.

Dmitry Davidov and Ari Rappoport. 2008b. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions. In *Proc. of ACL-HLT*, pages 692–700, Columbus, Ohio.

Dmitry Davidov, Ari Rappoport, and Moshe Koppel. 2007. Fully unsupervised discovery of concept-specific relationships by web mining. In *Proc. of ACL*, pages 232–239, Prague, Czech Republic.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010a. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proc. of CoNLL*, pages 107–116, Uppsala, Sweden.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010b. Enhanced sentiment learning using twitter hashtags and smileys. In *Proc. of Coling*, pages 241–249, Beijing, China.

Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence*, 19(1-2):109–123.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proc. of the 4th Web as Corpus Workshop*, WAC-4.

Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Carole E Chaski. 2007. Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *Int Journal of Digital Evidence*, 6(1):1–18.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of Coling – Volume 2*, pages 539–545, Stroudsburg, PA, USA.

Johan F Hoorn, Stefan L Frank, Wojtek Kowalczyk, and Floor van der Ham. 1999. Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, 14(3):311–338.

Shunichi Ishihara. 2011. A forensic authorship classification in sms messages: A likelihood ratio based approach using n-gram. In *Proc. of the Australasian Language Technology Association Workshop 2011*, pages 47–56, Canberra, Australia.

Patrick Juola. 2012. Large-scale experiments in authorship attribution. *English Studies*, 93(3):275–283.

Bradley Kjell, W Addison Woods, and Ophir Frieder. 1995. Information retrieval using letter tuples with neural network and nearest neighbor classifiers. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1222–1226. IEEE.

Bradley Kjell. 1994. Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9(2):119–124.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proc. of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, page 72.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proc. of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 624–628, New York, NY, USA.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *SIGIR*, pages 659–660.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *JMLR*, 8:1261–1276.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):9–26.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011a. Unsupervised decomposition of a document into authorial components. In *Proc. of ACL-HLT*, pages 1356–1364, Portland, Oregon, USA.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011b. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym

pattern linkage graphs. In *Proc. of ACL-HLT*, pages 1048–1056, Columbus, Ohio.

Robert Layton, Paul Watters, and Richard Dazeley. 2010. Authorship attribution for twitter in 140 characters or less. In *Proc. of the 2010 Second Cybercrime and Trustworthy Computing Workshop*, CTC '10, pages 1–8, Washington, DC, USA. IEEE Computer Society.

Robert AJ Matthews and Thomas VN Merriam. 1993. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. *Literary and Linguistic Computing*, 8(4):203–209.

DL Mealand. 1995. Correspondence analysis of luke. *Literary and linguistic computing*, 10(3):171–182.

Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science*, ns-9(214S):237–246.

George K Mikros and Kostas Perifanos. 2013. Authorship attribution in greek tweets using authors multilevel n-gram profiles. In *2013 AAAI Spring Symposium Series*.

Ashwin Mohan, Ibrahim M Baggili, and Marcus K Rogers. 2010. Authorship attribution of sms messages using an n-grams approach. Technical report, CERIAS Tech Report 2011.

Frederick Mosteller and David Lee Wallace. 1964. *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Fuchun Peng, Dale Schuurmans, and Shaojun Wang. 2004. Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317–345.

Conrad Sanderson and Simon Guenter. 2006. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proc. of EMNLP*, pages 482–491, Sydney, Australia.

Rui Sousa Silva, Gustavo Laboreiro, Luís Sarmento, Tim Grant, Eugénio Oliveira, and Belinda Maia. 2011. 'twazn me!!! ;(' automatic authorship analysis of micro-blogging messages. In *Proc. of the 16th international conference on Natural language processing and information systems*, NLDB'11, pages 161–168, Berlin, Heidelberg. Springer-Verlag.

Thamar Solorio, Sangita Pillay, Sindhu Raghavan, and Manuel Montes-Gomez. 2011. Modality specific meta features for authorship attribution in web forum posts. In *Proc. of IJCNLP*, pages 156–164, Chiang Mai, Thailand, November.

Efstathios Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. *Inf. Process. Manage.*, 44(2):790–799.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm–a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proc. of ICWSM*.

Peter Turney. 2008a. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proc. of Coling*, pages 905–912, Manchester, UK, August. Coling 2008 Organizing Committee.

Peter D. Turney. 2008b. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proc. of Coling*, pages 1–7, Stroudsburg, PA, USA.

George Udny Yule. 1939. On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika*, 30(3-4):363–390.