Applying Clustering Algorithms to Determine Authorship of Chinese Twitter Messages

Jinny Yan
Electrical Engineering & Computer Science
United States Military Academy
West Point, NY 10996
Email: jinnyyan@ccs.neu.edu

Suzanne J. Matthews
Electrical Engineering & Computer Science
United States Military Academy
West Point, NY 10996
Email: suzanne.matthews@usma.edu

***Abstract*—Author attribution research of character-based languages such as Chinese is still in its early stages. In this paper, we study the effectiveness of two popular clustering algorithms in determining the authorship of Chinese Twitter messages. We create a data-set of ten authors with 100 tweets each from publicly-available Chinese Twitter profiles. We analyze the data using simple k-means (SKM) and Expectation Maximization (EM), two popular clustering algorithms available in the Waikato Environment for Knowledge Analysis (WEKA). Our feature set includes character n-grams and Chinese function words derived from the literature. We achieve accuracy up to 44.53% for three authors, 29.24% for five authors, and 20.52% for ten authors. For our data-sets and the number of authors we compared, SKM returns better accuracy ratings. Lastly, we determine that function words are valuable features in attributing Chinese Tweets, and identify which of these Chinese function words were of most value.**

I. Introduction

The study of electronic authorship attribution has grown in recent years due to the prevalent use of the internet as a principle medium for communication. However, changing the language and type of electronic medium (articles, e-mails, blog posts etc.) often requires different analysis methods and utilized features. This is due to shifts in the manner these mediums are used, the different lengths, and the amount of meta-data surrounding the readable text. In character-based languages such as Chinese, there are over $50,000$ characters that an author could utilize. It is suggested that in order to read a full piece of elementary Chinese material, a person must know at least 500 commonly used Chinese characters. In comparison to Latin-based languages, this linguistic barrier increases the difficulty of quantifying and defining writing styles in Chinese. Thus, new features and methods must be explored in order to allow for effective authorship analysis of different electronic mediums in the Chinese language.

In this paper, we seek to study the efficacy of two popular clustering algorithms, simple k-means (SKM) and Expectation Maximization (EM), in identifying Chinese language authors in the Twitter-sphere. Twitter [1], a popular micro-blogging website, limits all users to 140 characters when writing posts. In this way, writing a Tweet requires a distinctively concise writing style for users to convey information. To the best of our knowledge, we are the first to study the authorship problem using Chinese Tweets.

We construct a Twitter corpus of ten Chinese authors, collecting 100 Tweets from each. We use established features of interest for the Chinese language, such as function words and character n-grams, to study how accurately each algorithm clusters Tweets by author. We were able to achieve up to 44.53% accuracy on three authors, 29.24% accuracy for five authors, and 20.52% for ten authors.

The rest of the paper is organized as follows. Section II discusses related work. Sections III and IV discusses our data and feature selection process. Our experimental process and results are described in Sections V and VI respectively. Finally, we discuss and conclude our work in Sections VII.

II. Related Work

Chinese authorship attribution is difficult for several reasons. In their research on character-level models, Peng *et. al.* [2] points to "word segmentation", a negative byproduct of Asian languages as a result of incorrectly grouped characters in an attempt to form meaning. Through analyzing semantic, syntactic, lexicographic, orthographic, and morphological linguistic devices in Greek, English, and Chinese, they determine that the most successful approaches to all character-based languages utilize an n-gram model [2]. The team uses popular modern Chinese novelists as their subjects of study, and conclude that there are $6,763$ most commonly used Chinese characters, $4,600$ of them being distinct. Using a 3-gram language model, they were able to achieve 94% accuracy [2].

Zheng *et. al.* [3] defines authorship attribution in a "multilingual context" and develops a feature set that includes several Chinese function words. These function words include terms such as "me" and "but" but also include those that do not translate easily to English, or are Chinese exclamative particles. Stamatatos *et. al.* [4] and Yu *et. al.* [5] note that the selection process for a function word list can be quite arbitrary. Stamatatos [4] notes that in deriving function word libraries for languages that may not have obvious function words, one option is to utilize the most frequently appearing characters. Yu [5] uses this strategy, deriving his own set of 35 Chinese function words from Jun Da's Modern Chinese Character Frequency List [6]. Yu [5] was able to achieve 90% accuracy on Chinese novels, 85% accuracy on Chinese essays, and 68% accuracy on Chinese language blog posts.

Twitter authorship is particularly difficult. Silva *et. al.* [7] suggests that standard features that work well for authorship analysis on larger texts (e.g. syntactic measures, lexical richness) fair poorly on extremely short texts. Ledger *et. al.* [8] indicates that 500 words is the minimum threshold to obtain good authorship results; Hirst *et. al.* [9] was able to achieve good authorship accuracy in texts of around 200 words. Silva [7] used quantitative and emotive markers to classify Portuguese Tweets derived from 40 sets of 3 authors. Schwartz *et. al.* [10] used n-grams to great effect on a collection of English tweets, achieving above 55% accuracy on 50 authors and 100 English tweets. Boutwell *et. al.* [11] achieves up to 40.3% accuracy on a data-set of 120 English tweets and 50 authors. Layton *et. al.* [12] notes that 120 tweets per user is an "important threshold" because anything past that number gave a "small but non-significant increase in accuracy" [12]. This research influenced us to limit our data collection to 100 tweets per author. However, note that the data-set discussed in [12] is made up of English Tweets.

While both English Twitter data and Chinese writing styles have been separately studied extensively in the past, we believe ourselves to be the first to explore the authorship of Chinese tweets.

III. Data Collection

We encountered several challenges in our data collection process. First, Twitter is blocked by the People's Republic of China (PRC). This severely limits our testing population to Twitter users in non-mainland territories such as Taiwan and Hong Kong, users in China that bypass the legal Firewall system, or Chinese-speaking users that immigrated to other countries.

Users in Taiwan or Hong Kong often include terms and vernacular choices in their Tweets that make it clearly identifiable that they are from those locations. Users in non-Chinese areas generally choose to write in the Chinese language with a substantial blend of the language of the country of residence, resulting in Tweets that lack enough usable Chinese material to analyze.

In the Peoples Republic of China (PRC), there exist three primary means of censorship: (1) "The Great Firewall of China", which prevents Chinese internet users from accessing certain Web sites altogether, (2) "keyword blocking", which sifts through an immense amount of data to prevent posts containing certain banned key words or phrases, and (3) "hand censoring", which differs greatly from the first two automated approaches of censorship [13]. Hand censoring is a labor-intensive process that requires employing many individuals to tend to the constantly changing, infinitely complex, and endlessly growing cyberspace. This process also becomes difficult to standardize. To meet these demands, the PRC government is believed to have developed automated means to identify rising problems in social media and to "[clip] social ties whenever any collective movements are in evidence or expected" [13].

While a large existing population of Twitter users residing in the PRC utilize a Virtual Private Network (VPN) to gain illegal access to Twitter, many of these users would choose to use Twitter as a compilation tool, sharing inspirational quotes, funny jokes, and occasionally, be a forum for confessions from a large body of anonymous Internet users. This creates a significant problem, especially when it isn't particularly obvious. If the Tweets were not written by a single user, this would impede the accuracy of identifying an author's unique style.

In addition to these inherent challenges, the PRC government is known to implant false profiles on Twitter that display resounding support for the government [13]. In this way, it is as though the government is "catphishing", or using a false online identity in interaction, the general public by promoting certain government activities. These mass-generated pieces of data have the potential to invalidate our findings. The potential ability to capture the exact features, algorithm, and settings that can properly model a Chinese Twitter users writing style is particularly frightening. If writing style can be modeled, this implies that it can also be forged. With such capabilities in hand, the Chinese government can increase its ability to censor its people.

We develop new Chinese Twitter corpora by utilizing a data collection tool called Forsight, developed by Crimson Hexagon [14]. Upon entering a set of keywords, Forsight is able to extract all of the related data and export it to a Microsoft Excel file.

Our first use for Forsight was to identify authentic authors from which to extract data. By inputting Chinese keywords that were used much more often in colloquial language, equivalent to English words like "he", "she", "they", "or", and "and", we were able to view a list of Chinese Twitter users that met a certain standard of word usage.

In order to gather Tweets that best represent an authors colloquial language, we used the native Forsight filter to eliminate Tweets that were retweets (Tweets directly taken from other users), even if a portion of the Tweet was written by the current author. To enable this, we chose prolific users that Tweet many times over the course of a day and have a history of thousands of Tweets.

We then searched through the list of authors that met the criteria and eliminated sources that were clearly incomplete, developing a list of ten authors. In order for Forsight to track a specific author, it creates "monitors" that specifically track an individual Twitter handle. We created ten monitors to extract 100 Tweets per author. Of our total ten authors, two were identified as women, and five as men. Three users had unknown genders.

IV. Feature Selection

We concentrated on function words and character n-grams for our experiments. In Chinese, a character could have a multitude of different functions and contextual meanings. Chinese "words" that have meaning in isolation can be found in a variety of forms: single characters, pairs, or even up to

Fig. 1. Function Character Library

Num. Authors	Feature Set	SKM	EM
3	Full	**44.53**	38.33
3	2-gram	37.63	38.90
3	3-gram	35.67	35.51
3	Function	43.31	**40.28**
5	Full	**29.24**	**27.90**
5	2-gram	24.36	22.81
5	3-gram	21.99	22.58
5	Function	28.72	25.85
10	Full	**20.52**	**17.23**
10	2-gram	14.94	10.36
10	3-gram	12.55	10.36
10	Function	19.32	17.03

TABLE I
AVERAGE PERCENT ACCURACY RETURNED FROM WEKA.

four-character idioms. We also considered the importance of Chinese function "phrases", function words that had more than one character, and decided to include character n-grams as features in our study.

A. Function Words

We decided early in our experiment that in order for us to create a finite-length library, we would limit our function word library to include only single character words. In choosing function words, it is important that they would be common enough and distinct enough to define a voice for each of the authors. For our function word library, we used the single character function words identified by Zheng [3] and Yu [5], shown in in Figure 1. Cells that are shaded black contain function words that were only derived from the set of function words from [5]. White cells contain function characters that were identified only in the study by [3]. Cells shaded gray contain function characters identified by both studies. Most of these characters have the ability to be utilized as conjunctions or pronouns, but could also be contextually modified into another form of speech.

Both Zheng [3] and Yu [5] used different strategies to create their function word libraries. Zheng [3] sought to find characters that (when used) contribute to an author voice. Yu [5] on the other hand, used commonly used characters that were usually used by all types of speakers in order to convey any sort of information. The discrepancy in function word selection between the two studies is unsurprising. Different libraries may emphasize particular forms of speech or have different sizes. In his paper, Stamatatos [4] notes that he has witnessed English function word sets that ranged from 150 to 675 words.

B. Character n-grams

We utilize two different variations of n-grams in our study. While it has been claimed in previous research that 3-grams are most effective for studying Chinese Authorship Attribution, we realize this may not apply to all writing mediums [2]. Tweets have much less text than novels, and this could have an impact on the ability to use n-grams to classify our Tweets.

In order to test this, we created two "sliding-window" functions for extracting 2-grams and 3-grams from our collection of Chinese Tweets. Much of the difficulty with executing this form of feature extraction was the @ and # symbols that would be embedded in the middle of the Tweets. This proved to be an obstacle, preventing a seamless transition through the piece of Unicode text.

V. Experimental Methodology

We used WEKA [15], a popular software package developed by researchers at the University of Waikato, to run our clustering algorithms. WEKA is one of the most popular packages for running data mining tasks. WEKA's user interface is simple and powerful, with various algorithms for classifying and clustering data.

For the scope of this project, we chose to focus on the simple k-means (SKM) and Expectation Maximization (EM) clustering algorithms in WEKA. Our goal was to determine how well our selected features could accurately cluster authors. Simple k-means and EM algorithms have the reputation of performing better than hierarchical clustering algorithms [16]. We were also motivated by the success of Iqbal *et. al.* [17] in using SKM and EM for successfully clustering authors.

Due to the aforementioned data restrictions, we collected only 100 Tweets per author. Our feature sets include 2-grams only (2-gram), 3-grams only (3-gram), function words only (Function), and all the features combined (Full). For each author, the top 10 most frequent features were considered. Modeling Iqbal *et. al.*'s experimentation, we chose to cluster samples of 3, 5 and 10 authors.

VI. Results

Table I provides an overview of our WEKA experiments. For each author number $n < 10$, we generated 10 random sets of n authors, and noted the accuracy of SKM and EM returned in WEKA. The numbers shown in Table I therefore represent the average of 10 runs.

Surprisingly, the 2-gram and 3-gram feature sets on their own did not provide the highest accuracies. For three authors, the Full set proved to be the most successful, yielding 44.53% accuracy with SKM. For our use of EM, the Function set was the most accurate, yielding a 40.28% accuracy with EM.

With five authors, the best result again came from SKM, which yielded a 29.24% accuracy when used on the Full set. The Full set was also the best result for EM which had an accuracy of 27.90%. Lastly, the best results came from SKM when we ran the algorithms on our entire collection of Tweets. SKM yields an accuracy of 20.52% on the Full set, while EM achieves an accuracy of 17.23%.

In all of the author sets (three, five, and ten authors), SKM outperformed EM in all categories using the Full set of features. We note the second highest accuracies are achieved when limiting the feature set to only function characters.

VII. Discussion & Conclusion

Our highest accuracy results are 44.53% for three authors, 29.24% accuracy for five authors, and 20.52% accuracy for ten authors. For all author sets, SKM had the highest overall accuracy numbers with the full feature set.

Our results suggest that function characters are very valuable for attributing authorship for Chinese Tweets. While the Full feature set produced the highest accuracy settings for each of our author sets, function characters alone produced higher accuracy percentages that 2-grams or 3-grams alone.

We theorize that in a character-limiting writing medium like Twitter, n-grams prove less effective because they require a large consumption of the character count. It is reasonable to expect that in shorter pieces of writing, writers choose to utilize common function characters to express themselves compared to a certain set of 2 or 3 characters. However, for larger numbers of authors, n-grams can certainly help distinguish between author writing styles.

We identify 10 function words that consistently yield high accuracy. Of these, three were part of the function word libraries of Zheng [3] and Yu [5]: "of", "the" and "yes". An additional three were included in the function library of [5]: "no/not", "in" and "have".

The ability to cluster authors in Chinese for Twitter means that there are undeveloped capabilities to effectively attribute authors in Chinese in any sort of written medium. There are many avenues of future research. First, we would like to expand our study to include users of Sina Weibo, a Twitter-like service that is popular in China.

We would also like to explore the efficacy of additional clustering algorithms for author attribution, such as DBSCAN [18], a density-based clustering approach. It can be particularly insightful and productive to look into newer, less-classic algorithms that may provide better results on higher-dimensional data [19].

We also would like to expand our Twitter corpora to a greater number of authors. We would like to to study how accuracy will change if we increase the number of Tweets per author beyond the 120 threshold. While Layton [12] advocates for this threshold, other authors have achieved good results with larger numbers of Tweets. While certain characteristics of studies on Tweets can be universal, the large language gap between Chinese and English leads us to leave room for error in applying Layton's [12] findings in our experiments.

Acknowledgment

We are grateful to Dr. Aaron Brantly and the Army Cyber Institute for enabling access to the Forsight Platform of Crimson Hexagon. This work was completed as part of Jinny Yan's computer science honors thesis during her undergraduate studies. The opinions in this work are solely of the authors and do not necessarily reflect that of the U.S. Army, the U.S. Military Academy, or the Department of Defense.

References

[1] Twitter, "Twitter [online]," website, last accessed, May 2016, https://twitter.com/.

[2] F. Peng, D. Schuurmans, S. Wang, and V. Keselj, "Language independent authorship attribution using character level language models," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 267–274.

[3] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[4] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[5] B. Yu, "Function words for chinese authorship attribution," in *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, 2012, pp. 45–53.

[6] J. Da, "Modern chinese character frequency list [online]," website, last accessed, May 2016, http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php.

[7] R. S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, and B. Maia, "twazn me!!!;(automatic authorship analysis of micro-blogging messages," in *Natural Language Processing and Information Systems*. Springer, 2011, pp. 161–168.

[8] G. Ledger and T. Merriam, "Shakespeare, fletcher, and the two noble kinsmen," *Literary and Linguistic Computing*, vol. 9, no. 3, pp. 235–248, 1994.

[9] G. Hirst and O. Feiguina, "Bigrams of syntactic labels for authorship discrimination of short texts," *Literary and Linguistic Computing*, vol. 22, no. 4, pp. 405–417, 2007.

[10] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel, "Authorship attribution of micro-messages," 2013.

[11] S. R. Boutwell, "Authorship attribution of short messages using multimodal features," Master's thesis, Naval Postgraduate School, 2011.

[12] R. Layton, P. Watters, and R. Dazeley, "Authorship attribution for twitter in 140 characters or less," in *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*. IEEE, 2010, pp. 1–8.

[13] G. King, J. Pan, and M. E. Roberts, "How censorship in china allows government criticism but silences collective expression," *American Political Science Review*, vol. 107, no. 02, pp. 326–343, 2013.

[14] C. Hexagon, "Forsight [online]," website, last accessed, May 2016, https://forsight.crimsonhexagon.com.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[16] O. A. Abbas *et al.*, "Comparisons between data clustering algorithms." *Int. Arab J. Inf. Technol.*, vol. 5, no. 3, pp. 320–325, 2008.

[17] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *digital investigation*, vol. 7, no. 1, pp. 56–64, 2010.

[18] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[19] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, p. 1, 2009.