WEKA in Forensic Authorship Analysis: A corpus-based approach of Saudi Authors

Mashael M. AlAmr, Professor Eric Atwell
School of English and School of Computing
University of Leeds, Leeds, United Kingdom
{enmmaa, e.s.atwell}@leeds.ac.uk

Abstract

This is a pilot study that aims to explore the potential of using WEKA in forensic authorship analysis. It is a corpus-based research using data from Twitter collected from thirteen authors from Riyadh, Saudi Arabia. It examines the performance of unbalanced and balanced data sets using different classifiers and parameters of word grams. The findings further support previous studies in computational authorship identification.

1 Introduction

"Authorship attribution, broadly defined, is one of the oldest and one of the newest problems in information retrieval." (Juola, 2008, p. 287). It aims to identify or attribute one or more disputed texts to a single or multiple author(s), either from a closed set or an open one (Stamatatos et al., 1999; Koppel et al., 2009). Recent trends in forensic authorship analysis aim for incorporating artificial intelligence tools to find reliable results that are free of cognitive biases. WEKA (Witten et al., 2016) is a collection of machine learning algorithms that perform data mining tasks. Its tools can achieve pre-processing, classification, clustering, and capable of developing new machine learning schemes. Therefore, WEKA is ideal to pre-process, classify, and even create machine learning schemes for identifying authorship. This study aims to explore the potential of using WEKA in a corpus-based forensic authorship analysis research.

1.1 Research questions

This study proposes to answer the following questions:

1- What is the size of data required for WEKA to identify authorship accurately?
2- Which classifier can accurately identify authorship using the NASCT corpus?
3- Which parameter is most accurate to identify authorship using the NASCT corpus?

2 Related Literature

This section discusses the notion of idiolect and related literature of artificial intelligence in authorship analysis using short texts.

2.1 The term 'idiolect'

This term was coined by Bloch (1948), which is a blend of the Greek words 'idio' and 'lect' to better reflect the concept of *personal language variety*. He defines idiolect as an individual-level variety that consists of a uniquely patterned set of linguistic characteristics. The notion of examining the individual's production of language was dismissed till a later stage in language studies. Crystal (1997) stated that each individual has their own language system that generates their unique dialect. Turell (2010) highlighted the importance of the concepts of markedness and saliency as art of idiolectal style in forensic text comparison. A text is a distinctive production thereby making it marked as it conveys specific and accurate information of its producer. The concept of saliency in linguistics is connected to the idea of a prominent feature that can be easily noticed. The concept of saliency that works best in forensic text comparison is a combined approach of discourse analysis and corpus linguistics. An item or a feature is considered salient if it stands out statistically when comparing two subcorpora or when a subcorpus is compared to the totality of a corpus (Turell, 2010). For this study, the linguistics features examined are dialectal features of Najdi Arabic, a dialect spoken in the central region of Najd where Riyadh, the capital of Saudi Arabia is

Proceedings of the 17th International Conference on Natural Language Processing, pages 257–260
Patna, India, December 18 - 21, 2020.

located. They are 45 dialect-specific features classified into interrogatives, negatives, and deictic expressions (Alothman, 2012; Binturki, 2015).

2.2 AI in Authorship analysis

Studies in authorship analysis, authorship identification in particular, aim to find the optimum classifiers, parameters, and n-grams that achieve the task with the highest accuracy rates. Numerous studies confirm that accurate authorship identification results can be achieved using small sized data (Rico-Sulayes, 2011; Brocardo et al., 2014; Saha et al. 2018). Moreover, several studies found that Linear Support Vector Machine (SVM) demonstrate accurate classifying results compared to others. Decision Tree J-48 and Multinominal Naïve Bayes perform more accurately with numeric data (Brocado et al., 2014; Maruktat et al., 2014).

In terms of n-grams, character grams proved to perform well in short texts such as WhatsApp and Twitter but with some limitations to identify authors' texts without cross examination (Shrestha et al., 2017; Banga et al., 2018). As for word grams, some studies conclude that unigrams even in the shape of an emoticon can show good results in indetifying authorship (Fissette, 2010). Bigrams proved to be successful in identifying authorship in literary texts (Feiguina and Hirst, 2007).

In addition, body of literature has been published in the authorship identification field in Saudi Arabia that focuses on computational approaches (Alruily, 2012; Althenayan and Menai, 2014; Al-Tuwairesh et al., 2015, 2018; Assiri et al., 2016) while linguistic and stylistic approaches fall short. This calls for a need to contribute to the field of forensic linguistics in general and forensic authorship analysis in Arabic in particular. Social media platforms such as Twitter are heavily populated by users who sometimes abuse such mediums. Saudis are responsible for 30% of the tweets posted (Salim, 2017). Simultaneously, there are efforts to fight cybercrime and issue regulations that incriminate hate speech and offensive language published online.

3 Methodology and corpus design

This section will demonstrate the NASCT corpus design, the data collection process, the sample selected for the study, and the pre-processing of the data and training WEKA.

3.1 Corpus design

Table 1 below shows the breakdown of the Najdi Arabic Specialized Corpus of Tweets.

Author	Tweets	Words
Faisal Alabdulkarim	2,825	60,213
Mansour Alrokibah	2,412	24,727
Abdulrahman Allahim	7,532	135,033
Ali Algofaily	2,424	37,217
Abdullah Alsubayel	5,805	35,791
Abdulaziz Alzamil	4,741	47,832
Taghreed Altassan	5,292	10,560
Wafa Alrasheed	2,200	24,236
Maha Alwabil	1,550	40,263
Arwa Almohanna	1,143	17,956
Ghadah Aleidi	7,952	70,709
Maha Alnuhait	2,089	27,784
Amani Alajlan	12,040	216,027
Total	58,005	748,348

Table 1: NASCT corpus design

3.2 Data collection

The data collected for this study are the authors' posts published on Twitter. To ensure authorship, all original tweets and replies were included in the corpus while retweets were eliminated. The corpus was compiled using Data Miner, a Google Chrome extension that identifies Arabic script. The time period of the data collection was March 1, 2018 – September 30, 2019. The data was produced as Excel sheets which the authors converted into .CSV file format.

3.3 Sample

The sample of the study are six males and seven females. All originated from the central region of Najd and current residents of Riyadh, the capital of Saudi Arabia. In terms of ethical considerations, all authors' accounts are public and verified thereby the tweets they post are public data. Lastly, all authors run the Twitter accounts personally.

3.4 Data preprocessing and training sets

To explore the NASCT using WEKA, the authors had to convert the .CSV files into ARFF file format. In order to train WEKA, the authors reassembled the corpus into thirteen separate ARFF files. They created two data training sets, the first is an unbalanced data set (TS1) which includes the full corpus. As shown in Table1, the subcorpora of some authors are substantially larger in size compared to others. The second one is a balanced data set (TS2), which includes equal number of tweets per author. Both data sets include 80% of the authors' data, and a header describing the types of linguistics attributes being examined. The remaining 20% of the authors' subcorpora was combined into one ARFF file for testing. Table 2 shows the number of tweets per data set.

Unbalanced data set = TS1	Balanced data set = TS2
44192	25247

Table 2: Training data sets

4 Findings and discussion

To examine different classifiers to see which performs most accurately, the authors ran a test using seven classifiers: Linear SVM, Multinominal Naïve Bayes, Decision tree J-48, KNN Depth=3, KNN Depth=5, Random Forest Estimator=5, and Random Forest Estimator=15. Table 3 shows the performance of seven different classifiers in three categories: unigrams, bigrams, and trigrams.

Parameter	Unigram	Bigrams	Trigrams
Linear SVM	0.59	0.6	0.6
M Naïve Bayes	0.47	0.48	0.48
J-48	0.4	0.4	0.4
KNN Depth=3	0.25	0.25	0.25
KNN Depth=5	0.25	0.25	0.25
Random FE=5	0.4	0.42	0.42
Random FE=15	0.46	0.47	0.47

Table 3: N-gram word models per classifier

The results of the first test show that Linear SVM scores the highest accuracy rates, therefore it was implemented in the next stage. The authors ran three tests to explore a range of parameters that can ensure the highest accuracy rates. In the first range, the minimum value is words that appear once and words that appear in 60% in the data files (min_df=1 – max_df=int (60/100)). The second one eliminates words that appear twice or less and words that occur in 80% of the data files (min_df=1 – max_df=int (80/100)). The last parameter test eliminates words that appear once and words that appear in 95% of the data files (min_df=1 – max_df=int (95/100)). Table 4 shows the accuracy rates of different parameters in both data sets.

Parameter	Unigram		Bigrams		Trigrams	
	TS1	TS2	TS1	TS2	TS1	TS2
1–60/100	0.59	0.58	0.6	0.6	0.59	0.6
2–80/100	0.59	0.58	0.6	0.45	0.6	0.29
0.001–95/100	0.49	0.58	0.49	0.59	0.49	0.59

Table 4: Accuracy rates per parameter using Linear SVM

In the first parameter test, both data sets scored the highest accuracy rates. The scores were most accurate across the three n-gram categories (0.59-0.6 respectively). The first data set TS1 scored a consistent and better performance compared to TS2 in the second parameter. The accuracy rates of the balanced data TS2 in the second parameter test were inconsistent. On the other hand, TS2 scored consistently higher results in the third parameter test compared to the unbalanced data set TS1. Nonetheless, both training data sets scored consistent results in unigrams, bigrams, and trigrams.

Furthermore, it appears that the balanced data set scores the highest overall results in unigrams, while the unbalanced data set scores the highest overall results in bigrams. However, the optimum parameter is the first test using bigrams.

5 Conclusion and future studies

This pilot study aimed to explore WEKA in forensic authorship analysis research. To answer the first question, the authors found that the unbalanced data set performed better than the smaller, balanced one. The large the size of the data provides WEKA with training and recognizing the features more accurately. As for which classifier performs best, results show that Linear SVM has the most accurate performance. This conforms to the findings of Fissette (2010) and Braocardo et al. (2014). Lastly, the results show that bigrams can accurately identify authorship, which confirms the findings of Feiguina and Hirst (2007).

For future studies, implementing different proportions for training and testing might yield higher, more accurate rates.

References

Alothman, E. 2012. *Digital Vernaculars: An Investigation of Najdi Arabic in Multilingual Synchronous Computer-Mediated Communication*. PhD thesis. University of Manchester.

Alruily, M. 2012: Saudi tweets dataset. figshare. Dataset.

Alshutayri, A and Atwell, E. 2018. Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers. In *Proceedings of OSACT'2018 Open-Source Arabic Corpora and Processing Tools. OSACT'2018 Open-Source Arabic Corpora and Processing Tools*, pages 07-12 May 2018, Miyazaki, Japan. (In Press)

Althenayan A. S. and Menai M.E-B. 2014. Naïve Bayes classifiers for authorship attribution of Arabic texts, *Journal of King Saud University - Computer and Information Sciences*, 26, pages 473-484.

Banga, R., Bhardwaj, A., Peng, S. L., & Shrivastava, G. 2018. Authorship Attribution for Online Social Media. In *Social Network Analytics for Contemporary Business Organizations*, pages 141-165. IGI Global.

Binturki, T. 2015. *The Acquisition of Negation in Najdi Arabic*. PhD thesis. University of Kansas.

Bloch, B. 1948. A set of postulates for phonemic analysis. *Language,* 24, pages 3-46.

Brocardo, M. L., Traore, I., and Woungang, I. 2014. Toward a framework for continuous authentication using stylometry. In *2014 IEEE 28th International Conference on Advanced Information Networking and Applications*, pages 106-115. IEEE.

Cotterill, J. 2010. How to use corpus linguistics in forensic linguistics. In O'Keeffe, A and McCarthy, M., eds., *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pages 66-79.

Crystal, D. 1997. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Feiguina, O. and Hirst, G. 2007. Authorship attribution for small texts: literary and forensic experiments. Paper presented to *the International Workshop on Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection*. 30th Annual International ACM SIGIR (SIGIR '07).

Fissette, M. 2010. Author identification in short texts.

Juola, P. 2008. Author Attribution, Foundations and Trends in Information Retrieval. (In Press)

Koester, A. 2010. Building small specialised corpora. In O'Keeffe, A and McCarthy, M., eds, *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pages 66-79.

Koppel, M., Schler, J. and Argamon, S. 2009. "Comp utational Methods in Authorship Attribution.". *Journal of the American Society for Information Science and Technology*, 60(no. 1), pages 9–26.

Rico-Sulayes, A. 2011. Statistical authorship attribution of Mexican drug trafficking online forum posts. *International Journal of Speech, Language & the Law*, *18*(1).

Saha, N., Das, P., & Saha, H. N. 2018. Authorship attribution of short texts using multi-layer perceptron. *International Journal of Applied Pattern Recognition*, *5*(3), pages 251-259.

Salim, F. 2017. *The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World* Dubai: MBR School of Government. Vol. 7.

Shrestha, P., Sierra, S., González, F. A., Montes, M., Rosso, P., & Solorio, T. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669-674.

Banga, R., Bhardwaj, A., Peng, S. L., & Shrivastava, G. 2018. Authorship Attribution for Online Social Media. In *Social Network Analytics for Contemporary Business Organizations*, pages 141-165. IGI Global.

Turell, M. T. 2010. The use of textual, grammatical, and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, 17(2), pages 211-250.

Twitter. About Twitter Verified Accounts. *URL: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts*.

WEKA.Witten, I. H., Frank, E., and Hall, M. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann Publishers, fourth edition.