Applying Authorship Analysis to Arabic Web Content

Ahmed Abbasi and Hsinchun Chen

Department of Management Information Systems, The University of Arizona,
Tucson, AZ 85721, USA
aabbasi@email.arizona.edu, hchen@bpa.arizona.edu

Abstract. The advent and rapid proliferation of internet communication has allowed the realization of numerous security issues. The anonymous nature of online mediums such as email, web sites, and forums provides an attractive communication method for criminal activity. Increased globalization and the boundless nature of the internet have further amplified these concerns due to the addition of a multilingual dimension. The world's social and political climate has caused Arabic to draw a great deal of attention. In this study we apply authorship identification techniques to Arabic web forum messages. Our research uses lexical, syntactic, structural, and content-specific writing style features for authorship identification. We address some of the problematic characteristics of Arabic in route to the development of an Arabic language model that provides a respectable level of classification accuracy for authorship discrimination. We also run experiments to evaluate the effectiveness of different feature types and classification techniques on our dataset.

1 Introduction

Increased online communication has spawned the need for greater analysis of web content. The use of potentially anonymous internet sources such as email, websites, and forums provides an attractive communication medium for criminal activity. The geographically boundless nature of the internet has further amplified these concerns due to the addition of a multilingual dimension. Application of authorship identification techniques across multilingual web content is important due to increased internet communication and globalization, and the ensuing security issues that are created.

Arabic is one of the official languages of the United Nations and is spoken by hundreds of million people. The language is gaining interest due to its socio-political importance and differences from Indo-European languages [10]. The morphological challenges pertaining to Arabic pose several critical problems for authorship identification techniques. These problems could be partially responsible for the lack of previous authorship analysis studies relating to Arabic.

In this paper we apply an existing framework for authorship identification to Arabic web forum messages. Techniques and features are incorporated to address the specific characteristics of Arabic, resulting in the creation of an Arabic language model.

The remainder of this paper is organized as follows: Section 2 surveys relevant authorship analysis studies, highlighting important writing attributes, classification

P. Kantor et al. (Eds.): ISI 2005, LNCS 3495, pp. 183 – 197, 2005.

techniques, and experiment parameters used in previous studies. Section 3 emphasizes the vital Arabic linguistic characteristics and challenges. Section 4 raises important research questions and describes an experiment designed to address these questions. Section 5 summarizes the experiment results and important findings. We conclude with a summary of key research contributions and point to future directions in Section 6.

2 Related Studies

In this section we briefly discuss previous research relating to the different writing style features, classification techniques, and experimental parameters used in authorship identification.

2.1 Authorship Identification

Authorship analysis is the process of evaluating writing characteristics in order to make inferences about authorship. It is rooted in the linguistic area known as Stylometry, which is defined as the statistical analysis of literary style. There are several categories of authorship analysis; however we are concerned with the branch known as authorship identification.

Authorship identification matches unidentified writings to an author based on writing style similarities between the author's known works and the unidentified piece. Studies can be traced back to the nineteenth century where the frequency of long words was used to differentiate between the works of Shakespeare, Marlowe, and Bacon [23]. Perhaps the most foundational work in the field was conducted by Mosteller and Wallace [24]. They used authorship identification techniques to correctly attribute the twelve disputed Federalist Papers.

Although authorship identification has its roots in historical literature such as Shakespeare and the Federalist papers, it has recently been applied to online material. De Vel et al. conducted a series of experiments on authorship identification of emails [8, 9]. Their studies provided an important foundation for the application of authorship identification techniques to the internet medium. Zheng et al. expanded de Vel et al.'s efforts by adding the multilingual dimension in their study of English and Chinese web forum messages [35]. In this study we are primarily interested in applying authorship identification to Arabic online messages.

2.2 Writing Style Features for Authorship Identification

Writing style features are characteristics that can be extracted from the text in order to facilitate authorship attribution. There are four important categories of features that have been used extensively in authorship identification; lexical, syntactic, structural, and content-specific.

Lexical features are the most traditional set of features used for authorship identification. They have their origins dating back to the nineteenth century, when Mendenhall used the frequency of long words for authorship identification in 1887

[23]. This set of features includes sentence length, vocabulary richness, word length distributions, usage frequency of individual letters, etc. [33, 34, 16].

Syntax is the patterns used for the formation of writing. Word usage (function words) and punctuation are two popular categories of syntactic features. Baayen et al. confirmed the importance of punctuation as an effective discriminator for authorship identification [3]. Mosteller and Wallace were the first to successfully use function words, which are generic words with more universal application (e.g., "while," "upon") [24].

Structural features deal with the organization and layout of the text. This set of features has been shown to be particularly important for online messages. De Vel et al. and Zheng et al. measured the usage of greetings and signatures in email messages as an important discriminator [8, 9, 35].

Content-specific features are words that are important within a specific topic domain. An example of content-specific words for a discussion on computer monitors might be "resolution" and "display." Martindale and McKenzie successfully applied content-specific words for identification of the disputed Federalist Papers [20].

A review of previous authorship analysis literature reveals a couple of important points. Firstly, lexical and syntactic features are the most frequently used categories of features due to their high discriminatory potential. Secondly, there is still a lack of consensus as to the best set of features for authorship identification. In a study done in 1998, Rudman determined that nearly 1,000 features have been used in authorship analysis [28].

2.3 Techniques for Authorship Identification

The two most commonly used analytical techniques for authorship identification are statistical and machine learning approaches. Several multivariate statistical approaches have been successfully applied in recent years. Burrows was the first to incorporate principle component analysis in 1987, which became popular due to its high discriminatory power [5]. Other successful multivariate methods include cluster analysis and discriminant analysis [15, 19].

Many potent machine learning techniques have been realized in recent years due to drastic increases in computational power. Tweedie et al. and Lowe and Mathews used neural networks whereas Diedrich et al. and de Vel et al. successfully applied SVM for authorship identification [30, 20, 8, 9]. Zheng et al. conducted a thorough study involving machine learning techniques in which they used decision trees, neural networks, and SVM [35].

Typically machine learning methods have achieved superior results as compared to statistical techniques. Machine learning approaches also benefit from less stringent requirements pertaining to models and assumptions. Most importantly, machine learning techniques are more tolerant to noise and can deal with a larger number of features [22].

2.4 Multilingual Authorship Identification

Applying authorship identification across different languages is becoming increasingly important due to the proliferation of the internet. Nevertheless, there has been a lack of studies focusing across different languages. Most previous studies have only focused on English, Greek, and Chinese. Stamamatos et al. applied authorship identification to a corpus of Greek newspaper articles [51]. Peng et al. conducted experiments on English documents, Chinese novels, and Greek newspapers [47]. Zheng et al. performed authorship identification on English and Chinese web forum messages [63]. In all previous studies, English results were better than other languages.

Applying authorship identification features across different languages is not without its difficulties. Since most writing style characteristics were designed for English, they may not always be applicable or relevant for other languages. Morphological and other linguistic differences can create feature extraction implementation difficulties. For example, Peng et al. noted that the lack of word segmentation in Chinese makes word-based lexical features (such as the number of words in a sentence) too difficult to extract [47]. They also found that the larger volume of words in Chinese makes vocabulary richness measures less effective.

2.5 Authorship Identification of Online Messages

Online messages present several problems for authorship identification as compared to conventional forms of writing (literary works, published articles). Perhaps the biggest concern is the shorter length of online messages. Ledger and Merriam claimed that authorship characteristics were less apparent below 500 words, while Forsyth and Holmes placed that number at 250 words [19, 13]. This problem is further amplified by the fact that online messages typically have a larger pool of potential authors to distinguish between. Most previous authorship identification studies performed on conventional writing involved 2-3 authors, with almost no studies exceeding 10 authors.

The less formal style of online messages can cause problems since there is a greater likelihood of misspelled words, use of abbreviations and acronyms (e.g., “j/k”), and unorthodox use of punctuations (e.g., “:)”). These differences can lead to inaccurate feature extraction. Such problems are more prevalent in online messages because authors are less likely to follow formal writing rules.

Despite all the challenges, the unique style of online messages may also provide helpful discriminators that are useful for identification. Structural features such as greetings, signatures, quotes, links, and use of phone numbers and email addresses as contact information can provide significant insight into an author’s writing characteristics.

3 Arabic Characteristics

Arabic is a Semitic language, meaning that it belongs to the group of Afro-Asian languages which also includes Hebrew. It is written from right to left with letters being joined together, similar to English cursive writing. Semitic languages have several characteristics that can cause problems for authorship analysis. These challenges include properties such as inflection, diacritics, word length, and elongation.

3.1 Inflection

Inflection is the derivation of stem words from a root. Although the root has a meaning, it is not a word but rather a class that contains stem instances (words). Stems are created by adding affixes (prefixes, infixes, and suffixes) to the root using specific patterns. Words with common roots are semantically related. Arabic roots are 3-5 letter consonant combinations with the majority being 3-letters. Al-Fedaghi and Al-Anzi believe that as many as 85% of Arabic words are derived from a tri-lateral root, suggesting that Arabic is highly inflectional [2]. Beesley estimated that there are approximately 5,000 roots in Arabic [4].

Figure 1 shows an inflection example. For the root and stems, the top row shows the word written using English alphabet characters and the second row shows the word written in Arabic. The words KTAB ("book") and MKTB ("office/desk") are derived from the root KTB. KTAB is created with the addition of the infix "A" whereas MKTB is derived with the addition of the prefix "M".

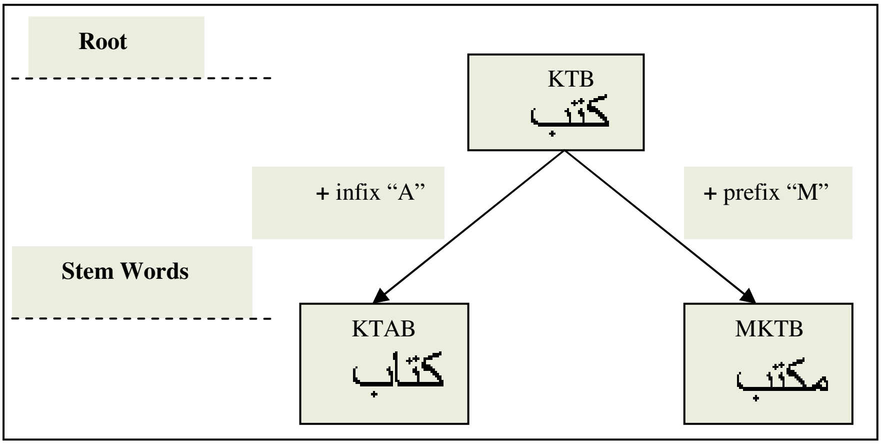

Fig. 1. Inflection Example

Larkey et al. stated that the orthographical and morphological properties of Arabic result in a great deal of lexical variation [18]. Inflection can cause feature extraction problems for lexical features because high levels of inflection increase the number of possible words, since a word can take on numerous forms. A larger word pool results in problems similar to those observed by Zheng et al. regarding Chinese [35]. In these circumstances vocabulary richness based measures such as Hapax, Lapax, Yule's, Honore's etc. will not be as effective.

3.2 Diacritics

Diacritics are markings above or below letters, used to indicate special phonetic values. An example of diacritics in English would be the little markings found on top of the letter "e" in the word résumé. These markings alter the pronunciation and meaning of the word. Arabic uses diacritics in every word to represent short vowels, consonant lengths, and relationships between words.

Although diacritics are an integral part of Arabic, they are rarely used in writing. Without diacritics, readers can use the sentence semantics to interpret the correct meaning of the word. For example, the words “resume” and “résumé” would look identical without diacritics, however a reader can figure out the appropriate word based on the context. Unfortunately this is not possible for computer feature extraction programs. This, coupled with the fact that Arabic uses diacritics in every word, poses major problems as illustrated by Figure 2.

English	Definitions	With Diacritics	Without Diacritics
MIN	who, whoever	مِن	من
MUN	from, of, for, than	مَن	من

Fig. 2. Diacritics Example

The Arabic function words MIN and MUN are written using identical letters, but with differing diacritics. MIN uses a short vowel represented by a marking underneath and “MUN” has one above. Without diacritics, the two words look identical and are indistinguishable for machines. This reduces the effectiveness of word-based syntactic features, specifically function words. From a feature extraction perspective, it is impossible to differentiate between the function word “who” (MIN) and “from” (MUN).

3.3 Word Length

Arabic words tend to be shorter than English words. The shorter length of Arabic words reduces the effectiveness of many lexical features. The short-word count feature; used to track words of length 3-letters or smaller, may have little discriminatory potential when applied to Arabic. Additionally, the word-length distribution may also be less effective since Arabic word length distributions have a smaller range.

3.4 Elongation

Arabic words are sometimes stretched out or elongated. This is done for purely stylistic reasons using a special Arabic character that resembles a dash (“-”). Elongation is possible because Arabic characters are joined during writing. Figure 3 shows an example of elongation. The word MZKR (“remind”) is elongated with the addition of four dashes between the “M” and the “Z” (denoted by a faint oval).

Although elongation provides an important authorship identification feature it can also create problems. Elongation can impact the accuracy of word length features. The example in Figure 3 causes the length of the word MZKR to double when elongated by fours dashes. How to handle elongation in terms of feature extraction is an important issue that must be resolved.

Elongated	English	Arabic	Word Length
No	MZKR	مذكر	4
Yes	M----ZKR	مـــذكر	8

Fig. 3. Elongation Example

4 Experiment

In this section we raise important research questions and discuss an experiment designed to address these questions. Specifically, we talk about the test bed, techniques, parameters, feature set, and experimental design used.

4.1 Research Questions

1. Will authorship analysis techniques be applicable in identifying authors in Arabic?
2. What are the effects of using different types of features in identifying authors?
3. Which classification techniques are appropriate for Arabic authorship analysis?

4.2 Experiment Test Bed

Our test bed consists of an Arabic dataset extracted from Yahoo groups. This dataset is composed of 20 authors and 20 messages per author. These authors discuss political ideologies and social issues in the Arab world.

4.3 Experiment Techniques

Based on previous studies, there are numerous classification techniques that can provide adequate performance. In this research, we adopted two machine learning classifiers; ID3 decision trees and Support Vector Machine.

ID3 is a decision tree building algorithm developed by Quinlan that uses a divide-and-conquer strategy based on an entropy measure for classification [27]. ID3 has been tested extensively and shown to rival other machine learning techniques in predictive power [6, 12]. Whereas the original algorithm was designed to deal with discrete values, the C4.5 algorithm extended ID3 to handle continuous values. Support Vector Machine (SVM) was developed by Vapnik on the premise of the Structural Risk Minimization principle derived from computational learning theory [32]. It has been used extensively in previous authorship identification studies [11, 35].

Both these techniques have been previously applied to authorship identification, with SVM typically outperforming ID3 [11, 35]. In this study we incorporated SVM for its classification power and robustness. SVM is able to handle hundreds and thousands of input values with great ease due to its ability to deal well with noisy data. ID3 was used for its efficiency. It is able to build classification models in a fraction of the time required by SVM.

4.4 Arabic Feature Set Issues

Before creating a feature set for Arabic we must address the challenges created by the characteristics of the language. In order to overcome the diacritics problem, we would need to embed a semantic-based engine into our feature extraction program. Since no such programs currently exist due to the arduous nature of the task, overcoming the lack of diacritics in the data set is not feasible. Thus, we will focus on the other challenges, specifically inflection, word length, and elongation.

4.4.1 Inflection

We decided to supplement our feature set by tracking usage frequencies of a select set of word roots. In addition to the inflection problem which impacts vocabulary richness measures, this will also help compensate for the loss in effectiveness of function words due to a lack of diacritics. Word roots have been shown to provide superior performance than normal Arabic words in information retrieval studies. Hmeidi et al. found that root indexing outperformed word indexing on both precision and recall in Arabic information retrieval [14]. Tracking root frequencies requires matching words to their appropriate roots. This can be accomplished using a similarity score based clustering algorithm.

4.4.1.1 Clustering Algorithm

Most clustering algorithms are intended to group words based on their similarities, rather than compare words to roots. These algorithms consist of several steps and some sort of equation used to evaluate similarity. The additional steps are necessary since word clustering is more challenging than word-root comparisons. Since we are comparing words against a list of roots, our primary concern is the use of a similarity-score based equation, and not necessarily all other parts of the algorithm. Two popular equations are Dice's equation and Jaccard's formula [1, 31]. Both these equations use n-grams to calculate the similarity score with the difference being that one places greater emphasis on shared n-grams that the other. The formulas are shown below:

SC(Dice) = 2 * (shared unique n-grams)/(sum of unique n-grams)

SC(Jac) = shared unique n-grams/(sum of unique n-grams – shared unique n-grams)

Although Dice's and Jaccard's equations are effective for English, they need to be supported by algorithms designed according to the characteristics of Arabic. De Roeck and Fares created a clustering algorithm based on Jaccard's formula, specifically designed for Arabic [7]. Their equation uses bi-grams, since they determined that bi-grams outperform other n-grams. The algorithm consists of five steps; however two steps require manual inspection and are not necessary for our purposes. Thus we omitted these parts and focused on Cross, Blank insertion, and applying Jaccard's formula.

Cross gives consonant letters greater weight by creating an additional bi-gram of the letter preceding and following a vowel. For example, in the word KTAB the letters before and after the vowel "A" form an additional bi-gram "TB", as shown in

Table 1. Word consonants are emphasized since roots are mostly composed of consonants, and giving consonants additional weight improves accuracy.

Table 1. Cross Example

Word	Without Cross	With Cross
KTAB	KT TA AB	KT TA TB AB

Blank insertion refers to the insertion of a blank character in the beginning and end of a word in order to give equal weight to border letters. For example, in the word KTAB without blank insertion the first and last letters are only used in a single bi-gram whereas the inner letters appear in 2 bi-grams each. Blank insertion allows all letters to be represented in an equal number of bi-grams, improving similarity score accuracy.

Table 2. Blank Insertion Example

Word	Without Blanks	With Blanks
KTAB	KT TA AB	*K KT TA AB B*

Table 3 shows a full example of our adaptation of De Roeck and Fares' clustering algorithm, using cross and blank insertion and applying Jaccard's formula. In this example, we compare the word KTAB against the roots KTB and KSB. The word is derived from the root KTB, and thus, it should receive higher similarity scores in comparisons with KTB as compared to KSB. In the comparison KTB scores 0.667 whereas KSB only gets a score of 0.25.

Table 3. Word-Root Comparison Example

Comparison	Total Unique bi-grams	Shared Unique bi-grams	Formula	SC
KSB KTAB	*K KS SB B* *K KA AT KT TB B*	*K B*	2/(10-2)	**0.25**
KTB KTAB	*K KT TB B* *K KA AT KT TB B*	*K KT B* KT	4/(10-4)	**0.67**

4.4.1.2 Applying the Word Root Feature

Root frequencies were extracted by calculating similarity scores for each word against a dictionary containing over 4,500 roots. The word was assigned to the root with the highest similarity score and the usage frequency of this root was incremented. Roots were sorted based on variance across authors. This was done based on the rationale that the roots with greater variance provide higher discriminatory potential. In order to determine the number of roots to include in the feature set, classification accuracy was used as the criteria. Between 0 and 50 roots were added to the complete Arabic feature set in order to determine the ideal quantity, which was tested using SVM as the classifier. Such a trial-and-error approach had to be used due to the lack of previous studies relating to Arabic authorship identification. Stamamatos et al. used a similar method to determine the ideal number of functions words to include in their study relating to Greek [29]. Figure 4 shows that the optimal number of roots using SVM was found to be 30.

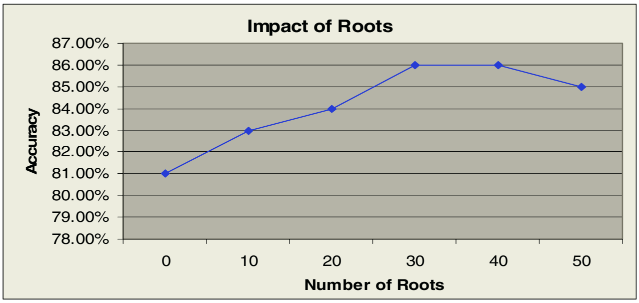

Fig. 4. Impact of Roots on Classification Accuracy

4.4.2 Word Length

Arabic words tend to be shorter than English words. Very long words are almost non-existent in Arabic. In order to test this hypothesis, we extracted the average length of our dataset and the number of words of length greater than 10. These values were then compared against the English dataset used by Zheng et al. which contained an equal number of web forum messages relating to computer software sales [35]. Table 4 shows that English words are approximately half a letter longer on average. More importantly, the number of English words longer than 10 letters is far greater. This disparity must be accounted for by tracking a smaller range for the Arabic word length distribution feature.

Table 4. English and Arabic Word Length Statistics

Data Set	Average Length	% Length > 10
English	5.17	6.125
Arabic	4.61	0.358

4.4.3 Elongation

The number of elongated words and the frequency of usage of elongation dashes should both be tracked since they represent important stylistic characteristics. Additionally, in order to keep word length features accurate, elongation dashes should not be included in word length measurements.

4.5 Arabic Feature Set

The Arabic feature set was modeled after the one used by Zheng et al. [35]. The feature set is composed of 410 features, including 78 lexical features, 292 syntactic features, 14 structural features and 11 content specific features, as shown in Table 5. In order to compensate for the lack of diacritics and inflection, a larger number of function words and 30 word roots were used. A smaller word length distribution and short word count threshold were also included. A larger set of content-specific words was incorporated due to the more general nature of the Arabic topics of discussion.

4.6 Experiment Procedure

Four feature sets were created. The first feature set contained only lexical features. The second set contained lexical and syntactic features. Structural features were added to the lexical and syntactic features in the third set. The fourth set consisted of all features (lexical, syntactic, structural, and content-specific). Such a step-wise addition of features was used for intuitive reasons. Past research has shown that lexical and syntactic features are the most important and hence, form the foundation for structural and content-specific features. In each experiment five authors were randomly selected and all 20 messages per author were tested using 30-fold cross validation with C4.5 and SVM.

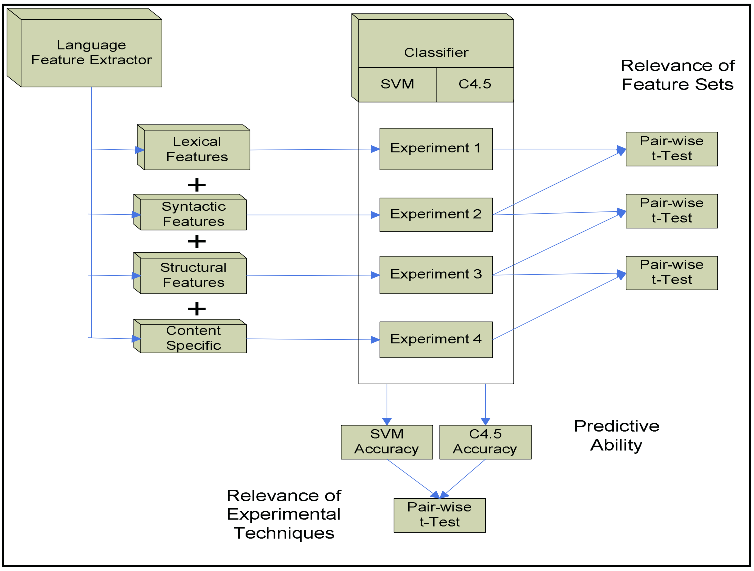

Fig. 5. Experiment Procedure

5 Results and Discussion

The results for the comparison of the different feature types and techniques are summarized in Table 6 and Figure 6. The accuracy kept increasing with the addition of more feature types. The maximum accuracy was achieved with the use of SVM and all feature types.

Table 6. Accuracy for Different Feature Sets across Techniques

	C4.5	**SVM**
F1	68.07%	74.20%
F1+F2	73.77%	77.53%
F1+F2+F3	76.23%	84.87%
F1+F2+F3+F4	81.03%	**85.43%**

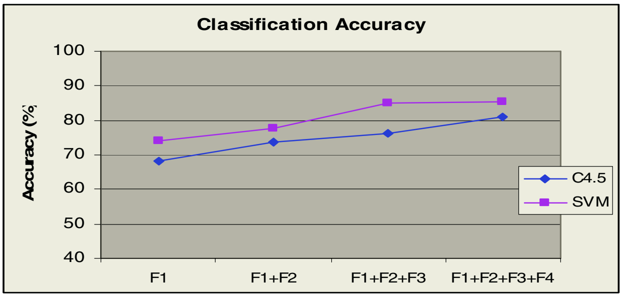

Fig. 6. Authorship Identification Accuracies for Different Feature Types

5.1 Comparison of Feature Types

Pairwise t-tests were conducted to show the significance of the additional feature types added. The results are shown in Table 7 below. An analysis of the t-tests and accuracies shows that all feature types significantly improved accuracy for the Arabic messages. Thus the impact of the different features types for Arabic was consistent with previous results found by Zheng et al. concerning online English and Chinese messages. The least effective set of features for Arabic was the content-specific words, however even this feature type had a significant impact with alpha set at 0.05. The lesser effectiveness of content-specific words could be attributed to the broad topic scope of the Arabic dataset.

Table 7. P-values of pairwise t-tests on accuracy using different feature types

t-Test Results with N=30		
Features/Techniques	C4.5	SVM
F1 vs F1+F2	**0.000*** **	**0.000*** **
F1+F2 vs F1+F2+F3	**0.000*** **	**0.000*** **
F1+F2+F3 vs F1+F2+F3+F4	**0.000*** **	**0.0134** **

**: significant with alpha = 0.05
***: significant with alpha = 0.01

5.2 Comparison of Classification Techniques

A comparison of C4.5 and SVM revealed that SVM significantly outperformed the decision tree classifier in all cases. This is consistent with previous studies that also showed SVM to be superior [11, 35]. The difference between the two classifiers was consistent with the addition of feature types, with SVM outperforming C4.5 by 4%-8%.

Table 8. P-values of pairwise t-tests on accuracy using different classifier techniques

t-Test Results with N=30				
Technique/Features	F1	F1+F2	F1+F2+F3	F1+F2+F3+F4
C4.5 vs SVM	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **

***: significant with alpha = 0.01

6 Conclusion and Future Directions

In this research we applied authorship identification techniques for the classification of Arabic web forum messages. In order to accomplish this we used techniques and features to overcome the challenges created by the morphological characteristics of Arabic. All feature types (lexical, syntactic, structural, and content-specific) provided significant discriminating power for Arabic, resulting in respectable classification accuracy. SVM outperformed C4.5 and the overall accuracy for Arabic was lower than previous English performance, both results being consistent with previous studies.

In the future we would like to analyze the differences between the English and Arabic language models using by evaluating the key features, as determined by decision trees. Emphasizing the linguistic differences between Arabic and English could provide further insight into possible methods for improving the performance of authorship identification methodologies in an online, multilingual setting.

Acknowledgements

This research has been supported in part by the following grant: NSF/ITR, "COPLINK Center for Intelligence and Security Informatics – A Crime Data Mining Approach to Developing Border Safe Research," EIA-0326348, September 2003-August 2005.

References

1. Adamson, George W. and J. Boreham (1974) The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information Storage and Retrieval,. Vol 10, pp 253-260
2. Al-Fedaghi Sabah S. and Fawaz Al-Anzi (1989) A new algorithm to generate Arabic root-pattern forms. Proceedings of the 11th National Computer Conference, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia., pp04-07.
3. Baayen, H., Halteren, H. v., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. Paper presented at the In Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002).

4. Beesley, K.B. (1996) Arabic Finite-State Morphological Analysis and Generation. Proceedings of COLING-96, pp 89-94.
5. Burrows, J. F. (1987). Word patterns and story shapes: the statistical analysis of narrative style. Literary and Linguistic Computing, 2, 61 -67.
6. Chen, H., Shankaranarayanan, G., Iyer, A., & She, L. (1998). A machine learning approach to inductive query by examples: an experiment using relevance feedback, ID3, Genetic Algorithms, and Simulated Annealing. Journal of the American Society for Information Science, 49(8), 693-705.
7. De Roeck, A. N. and Al-Fares, W. (2000) A morphologically sensitive clustering algorithm for identifying Arabic roots. In Proceedings ACL-2000. Hong Kong, 2000.
8. De Vel, O. (2000). Mining E-mail authorship. Paper presented at the Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000).
9. De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining E-mail content for author identification forensics. SIGMOD Record, 30(4), 55-64.
10. Diab, Mona**,** Kadri Hacioglu and Daniel Jurafsky. Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. Proceedings of HLT-NAACL 2004
11. Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2000). Authorship attribution with Support Vector Machines. Applied Intelligence.
12. Dietterich, T.G., Hild, H., & Bakiri, G., (1990), A comparative study of ID3 and Backpropagation for English Text-to-Speech mapping, Machine Learning, 24-31.
13. Forsyth, R. S., & Holmes, D. I. (1996). Feature finding for text classification. Literary and Linguistic Computing, 11(4).
14. Hmeidi, I., Kanaan, G. and M. Evens (1997) Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. Journal of the American Society for Information Science, 48/10, pp 867-881.
15. Holmes, D. I. (1992). A stylometric analysis of Mormon Scripture and related texts. Journal of Royal Statistical Society, 155, 91-120.
16. Holmes, D. I. (1998). The evolution of stylometry in humanities. Literary and Linguistic Computing, 13(3), 111-117.
17. Hoorn, J. F., Frank, S. L., Kowalczyk, W., & Ham, F. V. D. (1999). Neural network identification of poets using letter sequences. Literary and Linguistic Computing, 14(3), 311-338.
18. Larkey, L. S. and Connell, M. E. Arabic information retrieval at UMass in TREC-10. In *TREC 2001*. Gaithersburg: NIST, 2001.
19. Ledger, G. R., & Merriam, T. V. N. (1994). Shakespeare, Fletcher, and the two Noble Kinsmen. Literary and Linguistic Computing, 9, 235-248.
20. Lowe, D., & Matthews, R. (1995). Shakespeare vs. Fletcher: a stylometric analysis by radial basis functions. Computers and the Humanities, 29, 449-461.
21. Martindale, C., & McKenzie, D. (1995). On the utility of content analysis in author attribution: The Federalist. Computer and the Humanities, 29(259-270).
22. Mealand, D. L. (1995). Correspondence analysis of Luke. Literary and Linguistic Computing, 10(171-182).
23. Mendenhall, T. C. (1887). The characteristic curves of composition. Science, 11(11), 237-249.
24. Mosteller, F., Frederick, & Wallace, D. L. (1964). Applied Bayesian and classical inference: the case of the Federalist papers (2 ed.): Springer-Verlag.
25. Mosteller, F., & Wallace, D. L. (1964). Inference and disputed authorship: the Federalist: Addison-Wesley.

26. Peng, F., Schuurmans, D., Keselj, V., & Wang, S. (2003). Automated authorship attribution with character level language models. Paper presented at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003).
27. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
28. Rudman, J. (1998). The state of authorship attribution studies: some problems and solutions. Computers and the Humanities, 31, 351-365.
29. Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35(2), 193-214.
30. Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural Network applications in stylometry: the Federalist papers. Computers and the Humanities, 30(1), 1-10.
31. Van Rijsbergen, C. J. Information retrieval. London: Butterworths, 1979.
32. Vapnik, V. (1995). The nature of statistical learning theory. New York: Springer Verlag.
33. Yule, G. U. (1938). On sentence length as a statistical characteristic of style in prose. Biometrika, 30.
34. Yule, G. U. (1944). The statistical study of literary vocabulary. Cambridge University Press.
35. Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2003). Authorship Analysis in Cybercrime Investigation. Paper presented at the In Proceedings of the first NSF/NIJ Symposium, ISI2003, Tucson, AZ, USA.