Applying Authorship Analysis to Extremist-Group Web Forum Messages

Ahmed Abbasi and Hsinchun Chen, *University of Arizona*

Evaluating the linguistic features of Web messages and comparing them to known writing styles offers the intelligence community a tool for identifying patterns of terrorist communication.

The speed, ubiquity, and potential anonymity of Internet media—email, Web sites, and Internet forums—make them ideal communication channels for militant groups and terrorist organizations. Analyzing Web content has therefore become increasingly important to the intelligence and security agencies that monitor these groups. Authorship analysis can assist this activity by automatically extracting linguistic features from online messages and evaluating stylistic details for patterns of terrorist communication. However, authorship analysis techniques are rooted in work with literary texts, which differ significantly from online communication.

Furthermore, the global nature of terrorist activity necessitates the analysis of multilingual content. Arabic has garnered specific attention in recent years for sociopolitical reasons that include possible ties between certain Middle Eastern groups and terrorism. Arabic has morphological characteristics that pose several critical problems to current authorship analysis techniques.

To explore these problems, we modified an existing framework for analyzing online authorship and applied it to Arabic and English Web forum messages associated with known extremist groups.[1] We developed a special multilingual model—the set of algorithms and related features—to identify Arabic messages, gearing this model toward the language's unique characteristics. Furthermore, we incorporated a complex message extraction component to allow the use of a more comprehensive set of features tailored specifically toward online messages. A series of experiments evaluating the models indicated a high level of success in identifying communication patterns.

Authorship analysis

Stylometry is a linguistic discipline that applies statistical analysis to literary style. It is the basis for authorship analysis, which evaluates writing characteristics to make inferences about who wrote it. There are two major approaches to authorship analysis:

- *Authorship identification* deals with attributing authorship of unidentified writing on the basis of stylistic similarities between the author's known works and the unidentified piece; it deals with classification problems.
- A*uthorship characterization* attempts to formulate an author profile by making inferences about gender, education, and cultural backgrounds on the basis of writing style.

Here we're primarily concerned with applying authorship identification to English and Arabic online messages. These efforts are part of the Dark Web project, a research initiative to identify and evaluate individuals and groups that use the umbrella of online anonymity to support extremist and terrorist activities. Our work centers around collecting relevant content and then performing a multifaceted analysis of these groups in order to paint a picture that can aid the law enforcement and research community in better understanding them. This process begins by determining the relevant language features and techniques to use. Unfortunately, the authorship analysis literature[2] lacks consensus on these topics even for traditional written communication. Our task is complicated by the requirements of new media and the Arabic language. For example, online messages are shorter and noisier and they have a greater num-

1541-1672/05/$20.00 © 2005 IEEE
Published by the IEEE Computer Society

ber of potential authors. These characteristics impact the authorship identification parameters (features and techniques). Similarly, the linguistic complexities of Arabic elongation, inflection, and diacritics (all discussed later) create issues for the authorship identification features.

Writing style features

Writing style features that facilitate authorship attribution fall into four categories: lexical, syntactic, structural, and content-specific.

Lexical features can be either word- or character-based. Word-based lexical features include such characteristics as total number of words, words per sentence, word length distribution, and vocabulary richness. Vocabulary richness measures include the number of words that occur once (*hapax legomena*) and twice (*hapax dislegomena*), as well as several statistical measures defined by previous studies (see, for example, the work by George Yule[3]). Character-based lexical features include total number of characters, characters per sentence, characters per word, and the usage frequency of individual letters.

Syntax refers to the patterns used to form sentences. This category of features consists of the tools used to structure sentences, such as punctuation and function words. Example function words are *while* and *upon*. Usage patterns of function words can be effective features for authorship identification. For example, the difference between using the word *thus* or *hence* might seem subtle, but it can constitute a significant stylistic difference.

Structural features, which deal with the text's organization and layout, have proved particularly important in analyzing online messages.[4] Researchers traditionally focused on word structures such as greetings and signatures or on the number of paragraphs and average paragraph length. Although these features are important discriminators, they don't capture the additional information contained in online messages. For example, fonts, images, and links are not writing style features per se, but they provide important insight into a writer's online style. The use of various font sizes and colors requires a conscientious effort, making it a style marker. Similarly, embedded images and icons or links to different types of Web sites can reflect an author's technical prowess. Evaluating technical characteristics in terms of how images, hyperlinks, and audiovisual media are used isn't novel; researchers have applied it to Web sites for almost a decade.[5] Thus, we propose a new subcategory of structural features, called *technical structure,* to encompass font, hyperlink, and embedded image characteristics.

Content-specific features are words that are important within a specific topic domain. An example of content-specific words for a discussion on computers might be *RAM* and *laptop*. The rationale for content-specific words is similar to that of other word usage features but at a finer level of granularity.

Analysis techniques

Statistical and machine learning techniques constitute the two most common analytical approaches to authorship attribution. Many multivariate statistical approaches such as principal component analysis have shown a high level of accuracy.[6] However, these approaches also have some pitfalls, including the need for more stringent models and assumptions.

> Greetings, signatures, quotes, links, and the use of contact information (such as phone or email) can offer significant clues to authorship identification.

Machine learning techniques emerged from the drastic increases in computational power over the past several years. These techniques include support vector machines (SVMs), neural networks, and decision trees. They have gained wider acceptance in authorship analysis studies in recent years[1] because they provide greater scalability than statistical techniques for handling more features, and they're less susceptible to noisy data.[1] These benefits are important for working with online messages, which involve classification of many authors and a large feature set.

Online message complications

Conventional forms of writing pose fewer problems for authorship identification than online messages do. Writing style markers are far less visible for messages shorter than a few hundred words, making identification difficult or even impossible. The larger pool of potential authors in online attribution situations further amplifies the problem.

Additional difficulties relate to the casual style of online communication. Email and forum postings tend to be less formal than traditional writing, resulting in more misspellings and abbreviations, unorthodox structures, and improper use of punctuation. Consequently, applying authorship identification to Web content intrinsically involves a quagmire of noisy data.

Despite the challenges, online messages' unique structural characteristics can also provide helpful identification discriminators. Greetings, signatures, quotes, links, and the use of contact information (such as phone or email) can offer significant clues to authorship identification. We can further enhance this set of features by including technical structure features such as hyperlinks and embedded image characteristics.

Multilingual issues

Applying authorship identification across different languages is becoming more important with the Internet's rapid proliferation as a communication medium. The ramifications of terrorist organizations such as Al-Qaeda lend special urgency to analyzing Arabic in online communications. Nevertheless, little multilingual research exists. Excepting a few studies on Greek and Chinese,[1,7,8] most authorship identification research addresses English language features and identification techniques. For example, word-based lexical features (such as the number of words in a sentence) work well for English writing but not for Chinese, which doesn't segment words.[7] Additionally, the larger volume of words in Chinese makes vocabulary richness measures less effective.[1]

Arabic characteristics

Arabic poses some unique challenges with respect to the language's structural and stylistic properties. It is a Semitic language belonging to the Afro-Asian group. Semitic language characteristics that can complicate authorship analysis include inflection, diacritics, word length, and elongation.

Inflection

Arabic is a highly inflected language, which means that it builds its vocabulary primarily through the derivation of stem words from a root. Arabic has approximately 5,000 roots,

each of which is a three- to five-letter consonant combination.[9] Stems are created by adding affixes (such as prefixes) to the root. More than 85 percent of Arabic words are derived from roots, and words with common roots are semantically related.[10] The orthographical and morphological properties of Arabic result in significant lexical variation, because words can take on numerous forms.[11] Inflection creates feature extraction problems owing to the larger number of possible words, which weakens vocabulary richness measures.[1]

Figure 1 shows an inflection example demonstrating the derivation of two words (KTAB, meaning *book*, and MKTB, meaning *desk*) from the root KTB. For the root and stems, the top row shows the word written using English alphabet characters. The second row shows the word written in Arabic. Because Arabic letters are joined, making it difficult for non-Arabic readers to decipher individual letters, the third row shows the decomposed Arabic word in parentheses. KTAB and MKTB are created with the addition of the infix *A* and the prefix *M*, respectively.

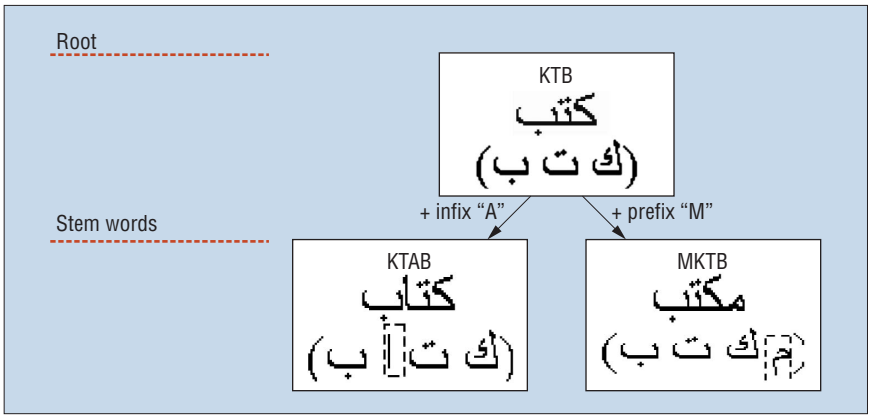

Figure 1. Inflection example. The decomposed letters for the Arabic words appear in parentheses to make it easier to see the infix and suffix stem words.

Table 1. Elongation example.

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

Diacritics

Diacritics are markings above or below letters used to indicate special phonetic values. In English, for example, a diacritic is the little mark on top of the letter *e* in the word *résumé*. These markings alter the word's pronunciation and meaning. Arabic uses diacritics in every word to represent short vowels, consonant lengths, and relationships between words; however, diacritics are rarely used in online communication. Although readers can use the sentence semantics to decipher proper meaning, this isn't feasible for an automated extraction program. For instance, without diacritics the words *resume* and *résumé* would look identical to a computer. The lack of diacritics can significantly impact the effectiveness of word-usage-based features such as function words. In Arabic, for example, it's impossible without diacritics to distinguish between the words *who* and *from*.

Word length and elongation

Arabic words are shorter than English words. This can reduce the effectiveness of many lexical features in identifying authorship. For example, word-length features are less discriminating because they are distributed over a smaller range. In addition, the use of longer English words is sometimes associated with greater writing complexity, but this assumption doesn't hold true for Arabic.

Elongation presents a further complication. Arabic words are sometimes elongated for purely stylistic reasons, using a special character that resembles a dash (—). Arabic characters are combined during writing, so elongation is possible by lengthening the joins between letters. Although it provides an important authorship style marker, elongation can also create problems. As table 1 illustrates, the word MZKR (*remind*) is extended with four short dashes between the *M* and the *Z* (denoted by a faint oval), doubling the word size. Thus, elongation can significantly inflate the values of word-length features. Handling elongation in terms of feature extraction is an important issue that must be addressed in our Arabic model.

Experiment design

We designed a series of experiments to test the efficacy of authorship identification techniques in an online setting. Our objective was to determine whether these techniques could identify authors writing in Arabic, how identification performance differed between English and Arabic, and the important feature differences between the English and Arabic groups and language models.

Test bed

Our test bed consisted of English and Arabic data sets extracted from Web forum messages. In both instances, we extracted 20 messages for each of 20 authors, resulting in a total of 400 messages per language. The average message length for the English data set was 76.6 words, and the average length for the Arabic data set was 580.69 words.

We derived the English messages from a US forum belonging to the White Knights, a chapter of the Ku Klux Klan. The KKK content revolved around political, racial, and religious issues. Members commonly used profanities and advocated the use of violence against groups they disliked. In addition to general anger and animosity, messages featured disturbing references to specific members of society. In some instances, the messages provided complete contact information, including street addresses, for these targeted individuals.

We extracted the Arabic data set from forum messages associated with the Palestinian Al-Aqsa Martyrs group. These strongly anti-America messages featured lengthy arguments espousing the group's views. The messages contained abundant embedded images and links relating to the war in Iraq and the treatment of Al-Qaeda prisoners. Authors used extremely graphic image content to support their central arguments. Much like the English message writers, authors in the Arabic forum advocated inflicting physical harm on groups they disliked.

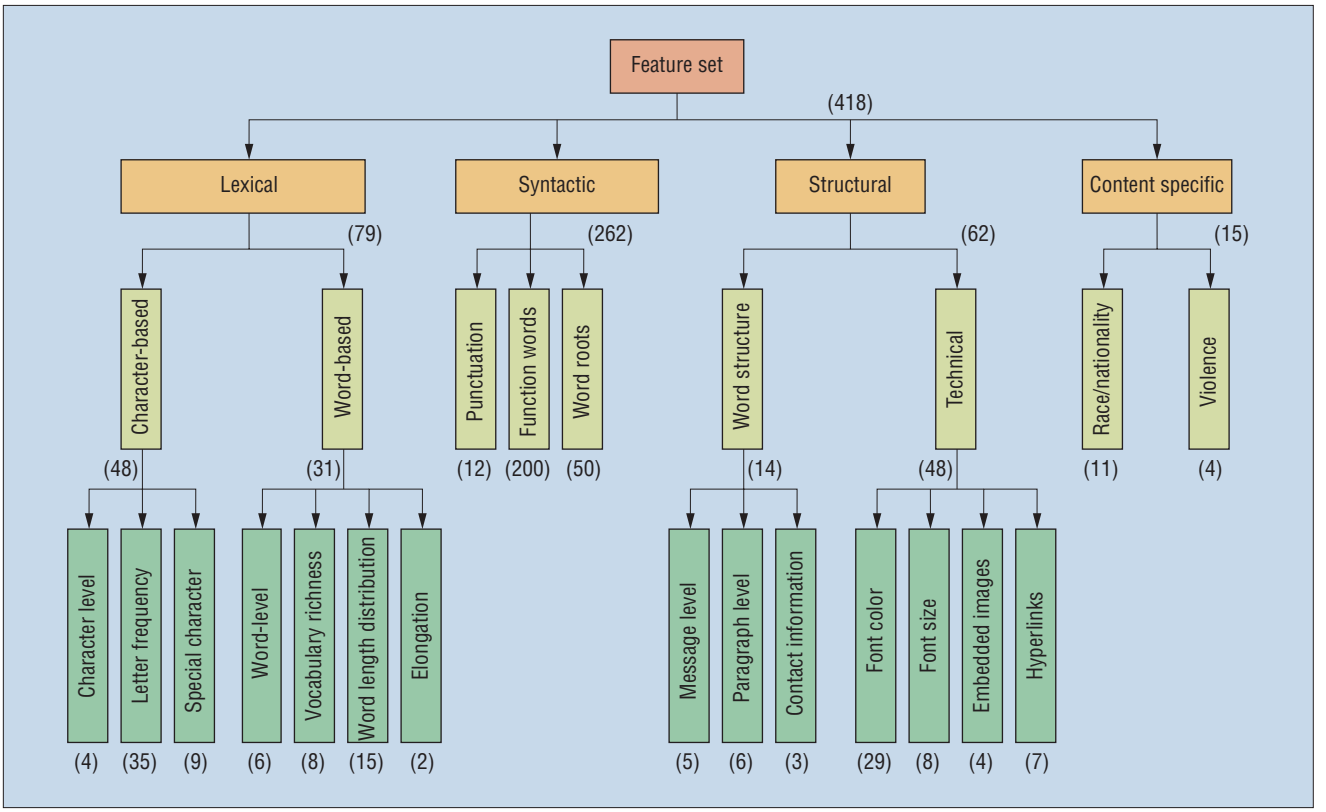

Figure 2. Arabic feature set. Modeled after the English set, it uses four feature types to characterize writing style: lexical, syntactic, structural, and content-specific.

Classifier techniques

In this study, we adopted two machine learning classifiers, C4.5 and SVM, applied in previous authorship analysis research.[1]

C4.5 is a powerful decision-tree-based classifier that rivals the performance of other machine learning techniques. We used C4.5 for its analytical and explanatory potential in effectively assessing key differences between the English and Arabic feature sets.

SVM is a computational learning method based on structural risk minimization. We incorporated SVM, which has gained popularity in recent years, for its classification power and robustness. SVM readily handles many input values owing to its capacity for dealing with noisy data.

Feature sets

We adapted an English feature set from previous online authorship studies.[1,4] The set consisted of 301 features, including 87 lexical, 158 syntactic, 45 structural, and 11 content-specific features. Our feature set differed mainly in the addition of four technical structure features: font color, font size, embedded images, and hyperlinks.

Figure 2 shows the Arabic feature set, which we modeled after the English set. It consists of 418 features, including 79 lexical, 262 syntactic, 62 structural, and 15 content-specific features.

Addressing Arabic characteristics. To create an effective Arabic feature set, we had to address the language's morphological and orthographical properties. To overcome the diacritics problem would have required using a semantic tagger. Because no feasible tagger solutions exist, we decided to focus on the challenges posed by inflection and by word length and elongation.

In the case of inflection, Arabic's heavy inflection means that root indexing outperforms word indexing on both precision and recall.[12] Accordingly, we complemented our feature set by tracking usage frequencies for a select set of word roots. In this way, we intended to help compensate for the losses in vocabulary richness measures.

Tracking root frequencies required a method for matching words to their appropriate roots. We used a clustering algorithm for this purpose. Anne De Roeck and Walid Al-Fares created a clustering algorithm specifically designed for Arabic.[13] Consisting of five steps, their algorithm is meant to compare words against other words as opposed to roots. Comparing words against a list of roots is an easier task, so we used only three of the algorithm's five steps.

We extracted root frequencies by calculating similarity scores for each word against a dictionary containing more than 4,500 roots. We assigned words to the root with the highest similarity score and incremented the selected root's usage frequency. An important issue was to determine the number of roots to include in the final feature set. We used a trial-and-error approach, as other multilingual authorship studies have done, because previous research hasn't yielded more definitive techniques.[8] To determine the number of roots to include, we added between 0 and 500 of the most frequently occurring roots to the complete Arabic feature set. We tested the classification power of these roots with SVM and integrated the optimal number (50 roots) into the feature set.

With regard to word length and elongation, we wanted to preserve elongation as an

important authorship style marker—in both its frequency and its length. At the same time, we wanted to eliminate any distortion of word-length distributions, because words longer than 10 characters are less common in Arabic than in English. Accordingly, to capture word length precisely, we embedded a filter in the Arabic feature extractor that removed elongation after it had been tracked.

English and Arabic feature set differences. After inspecting the data sets, we found 15 different font colors in the English messages and more than 120 in the Arabic. A closer look showed that many Arabic font colors were minor modifications of standard colors, which inflated the count. Because most of these modified colors were seldom used, we opted to avoid overfitting by excluding them from the feature set. The consolidated color count ultimately consisted of 12 colors for English and 29 for Arabic. We also included eight font-size, four embedded-image, and seven hyperlink technical structure features.

Table 2 highlights the differences between the English and Arabic feature sets. To compensate for the lack of diacritics and inflection, we used many function words and 50 word roots. The Arabic data set also included a smaller word-length distribution and short-word threshold.

Table 2. Key differences between English and Arabic feature sets.

Feature type	Feature	English feature set	Arabic feature set
Lexical	Short word count	<= 3	<= 2
	Word length distribution	1–20	1–15
	No. of elongations	n/a	2
Syntactic	No. of function words	150	200
	No. of word roots	n/a	50
Structural	No. of technical structures	31	48

Identification process

The complete online authorship identification process consisted of three main steps: collection, extraction, and experimentation. Figure 3 shows the complete process design for Arabic authorship identification.

Collection and extraction

We used spidering programs to identify Web forums of interest. These programs crawled through the Internet searching for Dark Web material, which is content involving potentially dangerous or criminal activity that might relate to cybercrime and homeland security issues. Once the process recognized such forums, collection programs stored the messages in text and HTML format. Extraction programs then derived writing style characteristics identified in the feature sets from each message.

The Arabic feature extractor was a bit more complex than the English one because it needed to account for elongation and inflection. We integrated an elongation filter, clustering algorithm, and root dictionary into the Arabic extraction process.

Experiments

After extracting the feature values, we categorized them into four feature sets. The first set (F1) consisted of lexical features, and the second (F2) encompassed lexical and syntactic features. We added structural features to the first two groups in the third feature set (F3) and inserted content-specific features with the other three categories in the fourth set (F4). Set F4 consequently contained all features: lexical, syntactic, structural, and content-specific. Such a stepwise increment of features reflected our perceptions concerning the feature categories' order of

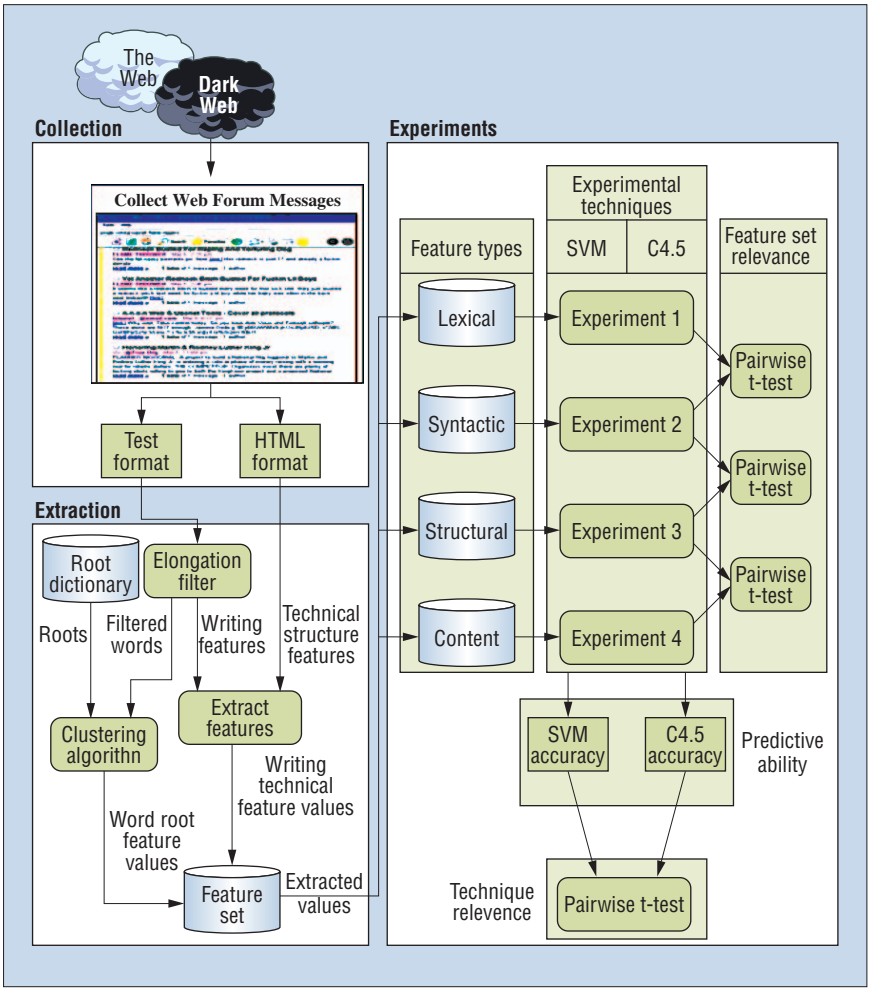

Figure 3. Authorship identification procedure for Arabic. From the Web data sets we collected (top left-hand corner), we extracted predefined language characteristics based on our models (bottom left-hand corner) and applied two feature classifiers—SVM and C4.5—to four experiments for determining the performance of authorship identification parameters (that is, features and techniques) (right-hand side).

Table 3. Accuracy for different feature sets using C4.5 and Support Vector Machine classifiers.

Feature sets	English data set		Arabic data set	
	Accuracy with C4.5 (%)	Accuracy with SVM (%)	Accuracy with C4.5 (%)	Accuracy with SVM (%)
F1	85.76	88.00	61.27	87.77
F1 + F2	87.23	90.77	65.40	91.00
F1 + F2 + F3	88.30	96.50	71.71	94.23
F1 + F2 + F3 + F4	90.10	97.00	71.93	94.83

Table 4. *P*-values of pairwise t-tests on classification accuracy using different feature types.

Features	t-test results for English data set, $N = 30$ 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*	0.1628
Features	**t-test results for Arabic data set, $N = 30$ 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	0.1216	0.0224†

*significant with alpha = 0.01
† significant with alpha = 0.05

importance. Studies have shown that lexical and syntactic features are the most important categories and hence form the foundation for structural and content-specific features.[1]

We applied this concept to test the relevance of feature categories for online English and Arabic messages. For the experiment, we created 30 randomly selected samples of five authors, which we used in all experiments. We evaluated each sample of five authors using all 20 messages per author and conducted a 30-fold cross validation with the C4.5 and SVM classifiers. The overall accuracy was the average precision (total number of correctly identified messages) across the 30 samples. We evaluated the feature type and classification accuracies using pairwise t-tests across the samples ($n = 30$).

Results and discussion

Table 3 summarizes authorship identification accuracy results for the comparison of the different feature types and techniques. The overall accuracies were exceptional, especially considering the task's difficulty and the results of previous authorship studies.[1,7,8] Perhaps most surprising was the relatively small drop in performance across languages. In both data sets, the accuracy kept improving with the addition of more feature types. We achieved maximum accuracy with the SVM classifier applied to all features for English and Arabic.

Feature type comparison

All feature categories improved classification accuracy in the stepwise analysis of features. Pairwise t-tests were conducted to show the statistical significance of the additional feature types added (for example, F2, F3, and F4). The results in table 4 show that all feature types significantly improved classification accuracy for Arabic and English, except for content-specific words. This feature category was statistically insignificant in two situations ($p = 0.1628$, $p = 0.1216$) and significant at a lower alpha-level in a third instance ($p = 0.0224$). The weaker performance of content-specific features could be attributable to their smaller representation in the feature set. The English and Arabic feature sets contained only 11 and 15 content-specific features, respectively. This number is far smaller than all other feature categories. Overall, the impact of the different feature types for Arabic was consistent with the results we obtained on English messages.

Classification technique comparison

Table 5 reveals that the SVM technique significantly outperformed the decision tree classifier in all cases. This is consistent with previous studies that have shown SVM to be better equipped to handle larger feature sets and noisier data, both characteristics that are associated with online authorship identification.[1,4] The difference in accuracy between classifiers across Arabic messages was far greater than it was in English messages: SVM outperformed C4.5 by more than 20 percent on all feature set combinations.

Analysis of English and Arabic group models

In evaluating the English and Arabic forum messages according to decision tree analysis and overall feature usage, we found key differences between the language models and some interesting trends pertaining to the two groups.

Decision tree analysis

The C4.5 decision tree is an effective analytical tool because of its descriptive nature. We can visualize decision trees to see the

Table 5. *P*-values of pairwise t-tests on accuracy using different classification techniques.

	t-test results for English data set, $N = 30$			
Technique/features	F1	F1 + F2	F1 + F2 + F3	F1 + F2 + F3 + F4
C4.5 vs. SVM	0.000*	0.000*	0.000*	0.000*
	t-test results for Arabic data set, $N = 30$			
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

Table 6. Decision-tree evaluation summary of key features.

	English messages			Arabic messages		
Features	No. of features used	Total no. of feature type in feature set	Percent used	No. of features used	Total no. of feature type in feature set	Percent used
Elongation	n/a	n/a	n/a	2	2	100
Word length	8	20	40	3	15	20
Punctuation	4	8	50	7	12	58.33
Function words	31	150	20.67	62	200	31
Root words	n/a	n/a	n/a	22	50	44
Word structure	8	14	57.14	8	14	57.14
Technical structure	12	31	38.71	32	48	66.67
Content-specific	3	11	27.77	3	15	20

effect of individual features, because trees choose the features with the highest discriminatory power, measured in terms of entropy reduction. We analyzed the C4.5 trees for the English and Arabic group models and extracted a list of the important features based on decision tree outputs.

Table 6 highlights the key differences between the English and Arabic models, according to the decision tree evaluations. The *percent used* column indicates the percentage of that feature group incorporated by the decision tree. The percentage provides a good basis for comparing the KKK and Al-Aqsa Martyr feature usage.

The features specifically integrated into the feature set to address the linguistic characteristics of Arabic played an important role based on the decision tree analysis. The C4.5 output showed that elongation features and nearly half the word roots were vital attributes that researchers should adopt in future studies. Furthermore, as expected, word length played a more critical role in the English KKK messages (40 percent) as compared to Arabic Al-Aqsa Martyr messages (20 percent).

The importance of punctuation, function words, and word-based structural features was fairly consistent across both languages. This result suggests that syntactical and structural characteristics are fairly robust feature categories across languages. The largest disparity in terms of feature importance was in the technical structure category. The use of font size, color, hyperlinks, and embedded images was more important in classifying messages from the Al-Aqsa Martyrs. The prevalence of technical structure features in the Arabic message group wasn't surprising—we expected a utilization of perhaps 50 percent—but the percentage used by the decision tree (66.7 percent) was surprising because it so exceeded our expectations.

Feature usage analysis

To provide a more in-depth analysis of the differences between the KKK and Al-Aqsa Martyr messages, we constructed a graph consisting of writing attributes common to the two groups. The visualization consisted of only lexical and structural features, because these feature groups are mostly language-independent. Figure 4 shows the average usage by language for each of these attributes.

We normalized the values to a 0-1 scale to facilitate more accurate comparisons. We identified five major feature groups within the lexical and structural categories: character-lexical, word-lexical, word-length, word-structure, and technical structure. We further decomposed these groups into subgroups (for example, paragraph structure) represented in either light gray or white in figure 4. In addition to demonstrating obvious linguistic dissimilarities, our comparison revealed several interesting subtleties that might be attributable to group or cultural differences.

Word/character lexical. The word- and character-level lexical features showed that the Al-Aqsa messages tended to be considerably longer than the KKK messages. In addition to overall length, sentence lengths of Al-Aqsa Martyr messages were longer, too.

Word length. Based on our data, mid-sized Arabic words in the 6-to-10-letter range were far more prevalent than English words of that length. However, longer Arabic words (greater than length 10) were less common. This is consistent with previous research[13] suggesting that Arabic has a narrower word-length distribution than English.

Word structure. Overall, the Al-Aqsa messages had a more formal structure, featuring more greetings, more sentences, and more—and lengthier—paragraphs. Unsurprisingly, author contact information was seldom provided, but the KKK authors more commonly supplied email addresses and phone numbers, which typically belonged to groups or individuals the author disliked.

Technical structure. Al-Aqsa messages used a plethora of font colors and sizes, often as tools to emphasize a certain point. Red, blue, and navy were almost as common as black. This was in sharp contrast to the KKK messages, where fonts featuring black, 10-to-12-point type were a fixture, with the exception of the occasional deviation to green or blue.

The Al-Aqsa messages had a far higher frequency of embedded images than the KKK messages (approximately 20 times more). Most of the disparity concerned GIF (graphics interchange format) and PNG (portable network graphics) file usage. The Al-Aqsa Martyr forum messages frequently used GIFs to represent slogans and logos; KKK messages used none. The Al-Aqsa group's messages also had many more links to static, dynamic, and image pages. Both forums used links to multimedia files; however, such direct links weren't common. Some multimedia links were provided via Web sites, so the parser classified them as Web page links.

Inferences. Both forums consisted of messages that stated opinions and beliefs. However, the structure and dynamics of the two groups' messages were noticeably different. The KKK forum messages were shorter and more conversational, implying greater familiarity between members. The Al-Aqsa group messages were more structured and formal,

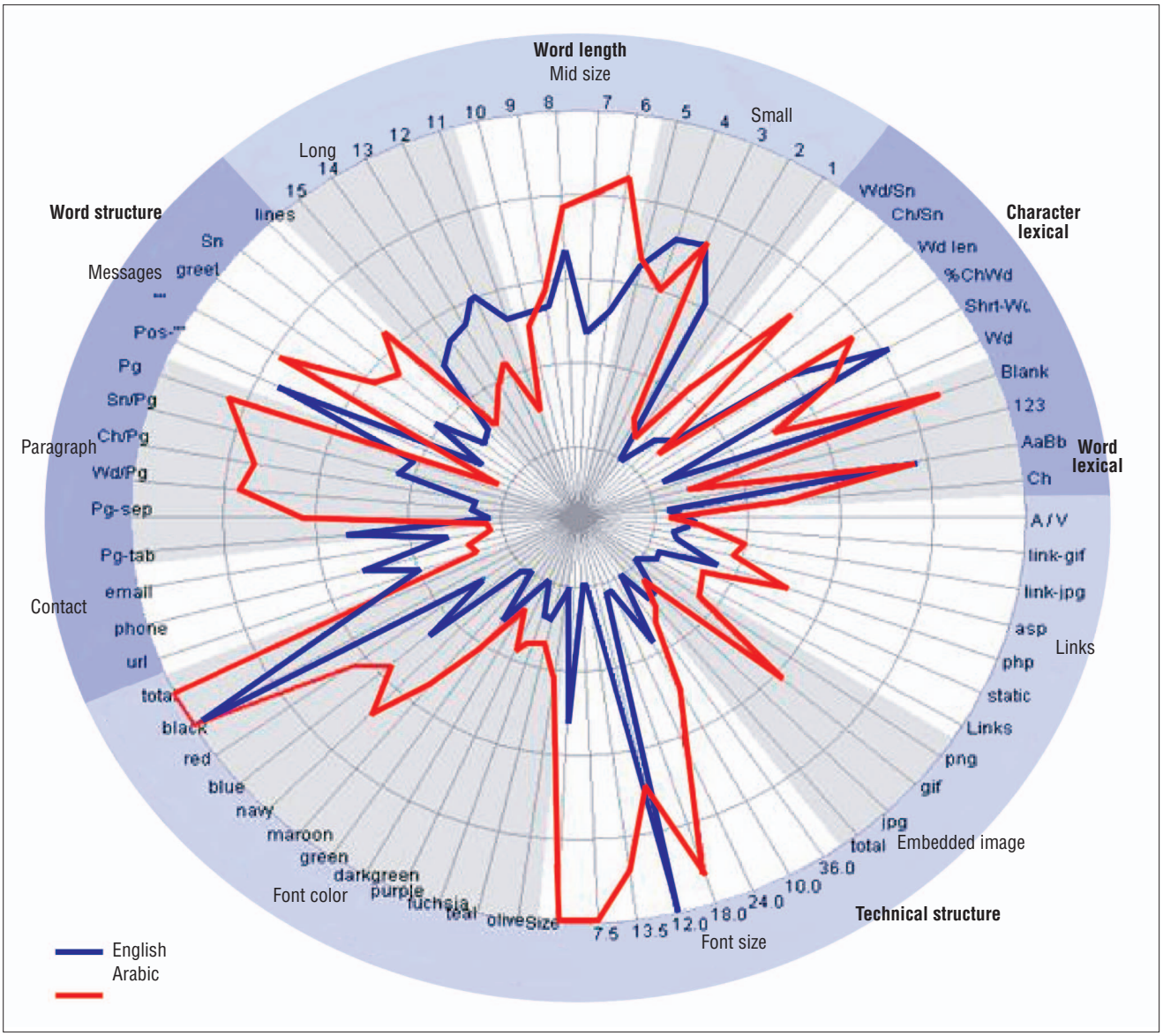

Figure 4. Comparison of group authorship characteristics. The shading differentiates feature sub-categories (such as short, medium, or long words).

and had a stronger persuasive inclination. The authors appeared to be making a concerted effort to state and justify their position by using a systematic, thorough writing approach. Bulleted points, paragraphs with headings, and generally longer message lengths, supported by embedded images and links, were the standard structural theme for the Al-Aqsa messages.

Our research showed significant discriminating power in the application of authorship identification techniques to English and Arabic extremist group forum messages. Having established a set of linguistic features and techniques for multilingual authorship analysis, we can pursue several potential future directions. The current authorship identification methodologies are limited in the number of authors we can apply them to. They require significant upward scalability to help discriminate between hundreds of potential authors. The development of more complex methodologies for differentiating between a larger set of authors is an important future endeavor.

We also plan a more comprehensive analysis of English and Arabic extremist group authorship tendencies to distinguish group-level differences from linguistic disparities inherent between English and Arabic. For example, do the "persuasive" tendencies observed regarding the Al-Aqsa Martyr messages have broader applicability to other extremist Arabic groups? Furthermore, what roles do geographic proximity and time play on group and individual authorship characteristics? Answers to these questions could prove of great value. ■

Acknowledgments

This research was supported by US National Science Foundation grant ITR-0326348, 2003-2005, "ITR: COPLINK Center for Intelligence and Secu-

The Authors

Ahmed Abbasi is a research associate at the Artificial Intelligence Laboratory and a doctoral student in the Management Information Systems department at the University of Arizona. His research interests include text mining, computer-mediated communication, information visualization, and knowledge management. Abbasi received his MBA from Virginia Tech in information systems. Contact him at the Artificial Intelligence Lab, Dept. of Management Information Systems, Univ. of Arizona, Tucson, AZ 85721; aabbasi@email.arizona.edu.

Hsinchun Chen is McClelland Professor of Management Information Systems at the University of Arizona. He received his PhD in information systems from New York University. Chen, whose most recent book is *Medical Informatics: Knowledge Management and Data Mining in Biomedicine* (Springer, 2005), also serves on several editorial boards including the *Journal of the American Society for Information Science and Technology*. Contact him at the Artificial Intelligence Lab, Dept. of Management Information Systems, Univ. of Arizona, Tucson, AZ 85721; hchen@eller.arizona.edu.

rity Informatics Research—A Crime Data Mining Approach to Developing Border Safe Research."

We are also grateful for the research assistance provided by fellow members of the Dark Web project team in the University of Arizona's Artificial Intelligence Laboratory, including Jialun Qin, Yilu Zhou, Greg Lai, and other team members who wish to remain anonymous.

References

1. R. Zheng et al., "A Framework of Authorship Identification for Online Messages: Writing Style Features and Classification Techniques," to be published in *J. Am. Soc. Information Science and Technology* (JASIST), 2005.
2. J. Rudman, "The State of Authorship Attribution Studies: Some Problems and Solutions, *Computers and the Humanities*, vol. 31, 1998, pp. 351–365.
3. G.U. Yule, *The Statistical Study of Literary Vocabulary*, Cambridge Univ. Press, 1944.
4. O. De Vel et al., "Mining E-mail Content for Author Identification Forensics," *SIGMOD Record*, vol. 30, no. 4, 2001, pp. 55–64.
5. J.W. Palmer and D.A. Griffith, "An Emerging Model of Web Site Design for Marketing," *Comm. ACM*, vol. 41, no. 3, 1998, pp. 44–51.
6. J.F. Burrows, "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style," *Literary and Linguistic Computing*, vol. 2, 1987, pp. 61–67.
7. F. Peng et al., "Automated Authorship Attribution with Character Level Language Models," presented at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003); http://users.cs.dal.ca/~vlado/papers/2003-EACL03-139.pdf.
8. E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-Based Authorship Attribution without Lexical Measures," *Computers and the Humanities*, vol. 35, no. 2, 2001, pp. 193–214.
9. K.B. Beesley, "Arabic Finite-State Morphological Analysis and Generation," *Proc. 16th Int'l Conf. Computational Linguistics* (COLING 96), 1996, Morgan Kaufmann, pp. 89–94.
10. S.S. Al-Fedaghi and F. Al-Anzi, "A New Algorithm to Generate Arabic Root-Pattern Forms," *Proc. 11th Nat'l Computer Conf.*, KFUPM, Saudi Arabia, 1989, pp. 391–400.
11. L.S. Larkey and M.E. Connell, "Arabic Information Retrieval at UMass in TREC-10," *Proc. 10th Text Retrieval Conf.* (TREC 2001), Nat'l Inst. of Standards and Technology, 2001.
12. I. Hmeidi, G. Kanaan, and M. Evens, "Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents," *J. Am. Soc. Information Science*, vol. 48, no. 10, 1997, pp. 867–881.
13. A.N. De Roeck and W. Al-Fares, "A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots," *Proc. Assoc. for Computational Linguistics* (ACL 00), 2000; www.informatik.uni-trier.de/~ley/db/conf/acl/acl2000.html.

For more information on this or any other computing topic, please visit our digital library at http://computer.org/publications/dlib.