Automatically Classifying Documents by Ideological and Organizational Affiliation

Moshe Koppel
Dept. of Computer Science.
Bar-Ilan University
Ramat-Gan, Israel
moishk@gmail.com

Navot Akiva
Dept. of Computer Science.
Bar-Ilan University
Ramat-Gan, Israel
navot.akiva@gmail.com

Eli Alshech
Middle East Media Research Institute
Jerusalem, Israel
eli_a@memri-tv.org

Kfir Bar
Intuview Corp.
Herzliya, Israel
kfir@intuview.com

Abstract—**We show how an Arabic language religious-political document can be automatically classified according to the ideological stream and organizational affiliation that it represents. Tests show that our methods achieve near-perfect accuracy.**

Keywords- Text Categorization, Author Profiling, Islamic Organizations

I. Introduction

In this paper we will present an automated system for classifying an Arabic language document according to ideological and organizational affiliation.

The term “ideological affiliation” refers to the doctrine underlying the documents. In particular, we consider four ideological streams: Salafi-Jihadi, Mainstream Islam (apolitical), Muslim Brotherhood, and Wahhabi. We selected these four because each is well known and has a significant following within contemporary Muslim societies and because each differs from the others in important doctrinal aspects.

The term “organizational affiliation” refers here to the religious group from which a particular document stems (e.g., the website of the Hamas). The organizations we consider are Hamas, Hizballah, Al Qaeda and Muslim Brotherhood. While organizational affiliation is often closely correlated with ideological affiliation, there is no strict mapping between the two. Thus, for example, the Muslim Brotherhood is a coherent socio-religious movement with defined religio-political institutions in Egypt and other countries. At the same time, however, it represents an ideology that is today adopted by many Islamic groups outside the Muslim Brotherhood Movement. In any case, we treat the two problems independently.

For security analysts and law enforcement agents, the ability to categorize documents by ideology or organization accurately and efficiently is of utmost importance. For example, it could help police ascertain a suspect's organizational affiliation based on documents found in his or her possession. In addition, such categorization allows officials to monitor and curb the proliferation in cyberspace of texts that are regarded as illegal or dangerous. While it is true that experts can categorize documents manually with little difficulty, applications such as these require the processing of thousands of documents in real time, a task that can only be accomplished using automated tools such as we propose here.

The outline of this paper is as follows. In the following section, we outline the central issues in automated text categorization, highlighting the differences between content-based and style-based text categorization as well as certain special considerations that arise with regard to Arabic documents. In Section 3, we describe our experimental framework and in Section 4, we provide results, focusing especially on the features that prove to be most useful for our tasks.

II. Text Categorization

In a text categorization [9] problem we are given two or more classes of documents and we need to find some classifier that automatically assigns a new document to the correct class, presumably based on statistical differences between documents in the respective classes. Thus, for example, we might wish to classify a document as being about one of a number of possible topics, as having been written by a certain type of author, as being spam or non-spam, as expressing positive or negative sentiment with regard to some issue, and so on.

Regardless of the particular dimension along which we wish to categorize documents, the first step, document representation, involves defining a set of text features which might potentially be useful for categorizing texts along that dimension and then representing each text as a vector in which entries represent (some non-decreasing function of) the frequency of each feature in the text. Once documents have been represented as vectors, some learning algorithm can be used to construct models that distinguish between vectors representing documents in the respective classes.

The two central questions that need to be addressed in designing a text categorization system are the choice of features to use and the choice of learning algorithm. With regard to the choice of features, it is perhaps useful to first consider the distinction between topic-based text categorization and style-based text categorization. When we wish to classify documents according to topic, it stands to reason that the features that are most likely to be helpful are simply content-bearing words. For example, documents about sports can be distinguished from documents about politics by checking the frequencies of sports-related or politics-related words. In contrast, for categorizing according to writing style, for example for purpose of authorship attribution, one needs to use precisely those linguistic feature types that are content-independent. Such features might include function words (content-free words like *and*, *the* and *if*), parts of speech, and a variety of other feature types [6]. What is interesting for our purposes now is that there are many text categorization problems – prominent examples include spam filtering and sentiment analysis – that are not obviously topic-based or style-based, but rather straddle both. As we shall see, the problems we consider in this paper, ideology and organizational affiliation, are also tied to both content and style.

Arabic texts present special problems in terms of feature selection for text categorization [1][2][3][8][10]. First of all, the same word roots can appear in myriad forms and it is non-trivial to cluster all these forms together. Moreover, function words tend to be conflated into word affixes in Arabic, thus decreasing the number of function words, but increasing the amount of morphological features that can be exploited. The richness of Arabic morphology also renders part-of-speech tagging a more difficult task in Arabic than in other languages, such as English, in which each part-of-speech is typically represented as a separate word.

We note that many of these issues are discussed in considerable detail in the important paper of Abbasi and Chen [1], the work most similar to ours. Abbasi and Chen sought to identify specific individual authors of documents posted on extremist sites. By contrast, our objective is to identify the organizations and ideologies to which these authors subscribe.

III. Experimental Design

In this section we discuss the corpora used for our experiments, the feature sets and learning methods used for finding classifiers, and the testing methodology used to evaluate results.

A. The Corpora

Our first corpus consists of a set of 552 documents each of which is manually labeled according to the organization that produced it: Hamas (150 documents), Al Qaeda (150 documents), Muslim Brotherhood (150 documents), and Hizballah (102 documents).

Our second corpus consists of 1485 documents manually labeled according to the ideological stream from which they originate: Salafi-jihadi (400 documents), Muslim Brotherhood (400 documents), Mainstream Islam (400 documents) and Wahhabi (285 documents). (Two clarifications are in order here. First, as mentioned above, Muslim Brotherhood is a defined socio-religious movement; at the same time, however, it represents an ideology adopted by many Islamic groups not affiliated with this movement. Second, Mainstream Islam represents an apolitical Islamic ideology.)

The documents in each corpus were taken from a wide variety of public sources to ensure representativeness. The documents range in length from 106 to 10,331 words with an average of 1253. Each document was cleansed of html and other non-text material, as well as tags and other identifying marks. Diacritics and kasheeda marks were also removed.

B. Extracting Features

Each document is represented as a numerical vector each entry of which is the frequency of some linguistic feature in the document. In our experiment, we simply chose as features the 1000 most common words in the entire corpus. We did not use stemming since this process is time-consuming and preliminary tests indicated that it is not necessary for achieving good results. Note that the total number of unique words appearing in the corpus is in the tens of thousands; we chose only the most frequently occurring 1000 words since it is well established [4] that filtering in this way increases efficiency without degrading accuracy. The 1000 words we chose include both function words and content words.

C. Learning Method

Given labeled vectors, we use some learning algorithm to build classifiers. In this study, we use Bayesian multi-class regression (BMR) [5], which has been shown to be extremely effective for text categorization [5][6]. One advantage of BMR is that it is a linear method so that each feature is assigned some weight in the classifier. This allows us to easily ascertain the most important features for each category.

D. Testing methodology

In order to test the accuracy of our methods on out-of-sample documents, we use ten-fold cross-validation: we randomly divide the corpus into ten sets, learn a classifier on nine sets and test on the held out set. We repeat this procedure ten times, each time holding out a different set.

IV. Results

A. Organizations

Using ten-fold cross-validation on the organization experiment with a feature set consisting of 1000 words, we find that *every one of the 552 test documents is correctly classified.*

Since we use a linear classifier, each feature is assigned a weight by the classifier for each category. This weight indicates the significance of that feature for identifying a document as being in that category. In Table I, we show the ten highest weighted words in each category. We find, for example, that the word *group* (الجماعة) is a marker of Muslim Brotherhood because the movement commonly refers to itself in texts as 'the group'. The word *Palestine* (فلسطين) is a marker of Hamas because it operates inside the Palestinian territories and its declared agenda is the conquest of Palestine. The prominence of the word *Unity* (التوحيد) in the documents of Al-Qaeda reflects the preoccupation with the doctrine of "Unity of God" in Al-Qaeda's writings. The word *Israel* is a marker for Hizballah since the other groups tend not to refer to Israel by name.

Less expectedly, we find a number of function words that serve as markers, such as *only* (إلا) in Al-Qaeda texts and *there* (هناك) in Hizballah texts. The phenomenon of function words such as these being useful for identifying distinct styles of writing is well attested [6][7], including specifically for Arabic documents [1]. Thus, in order to test the extent to which such style-based features are themselves adequate for distinguishing organizations, we ran the experiment using a feature set consisting only of 82 Arabic function words. Even for this small and apparently weak feature set, we obtain accuracy of 80%.

TABLE I. TEN HIGHEST WEIGHTED WORDS PER ORGANIZATION

Organization			
Al Qaeda	***Hamas***	***Hizballah***	***Muslim Brothers***
المجاهدين (Mujahideen)	اوكان (And Was)	لبنان (Lebanon)	عز (Ezz)
التوحيد (Unification)	وقد (and Already)	الإمام (Imam)	بيان (Communiqué)
قد (Already)	شعبنا (Our People)	حزب (Party)	له (To it)
إلا (Only)	الناصر (Nasser)	الى (To)	الإسلامي (Islamic)
أمريكا (America)	صلاح (Salah)	التي (Which)	مطلوب (Required)
الجهاد (Jihad)	الفلسطيني (Palestinian)	السيد (Mister)	المسلمين (Muslims)
القول (Statement)	العمليات (Operations)	بلدة (Town)	العمل (Work)
السلاح (Weapon)	فلسطين (Palestine)	هناك (There)	الاجتماعي (Social)
فإن (and if)	ألوية (Bridgades)	إسرائيل (Israel)	الجماعة (Group)
العدد (Number)	او (or)	اللبنانية (Lebanese)	فهو (It is)

B. Ideology

Using ten-fold cross-validation on the ideology experiment with a feature set consisting of 1000 words, we find that *all but 2 of 1485 test document are correctly classified*. The two exceptions are Wahhabi documents that were misclassified as Salafi-jihadi. Examination of the documents indicates that their content could easily have been assigned to either ideology.

In Table II, we show the ten highest weighted words in each category. We find, for example, that the word *resistance* (المقاومة) is a marker of Muslim Brotherhood documents referring to activities against foreign military presence in Muslim countries. By contrast, *jihad* (الجهاد) is a marker of Salafi-jihadi ideology since strict adherents to the Salafi-jihadi ideology prefer this word to *resistance*, which might suggest that waging war against the West is justified only to resist Western occupation rather than mandatory at all times for the purpose of subjecting the West to Islamic law. The word *Mecca* (مكة) is a marker of Wahhabi documents because the Wahhabi doctrine is the official creed of Saudi Arabia.

Here, too, we find a number of function words that are useful. For example, the word *was* (كان) is a marker of Mainstream Islam documents and *to* (الى) is a marker of Muslim Brotherhood ideology. In fact, even when using only function words as features, we obtain accuracy of 73%.

TABLE II. TEN HIGHEST WEIGHTED WORDS PER IDEOLOGY

Ideology			
Wahhabi	***Muslim Brothers***	***Mainstream***	***Salafi - Jihadi***
الحق (Truth)	الى (To)	سؤال (Question)	تال (Following)
قلت (I Said)	الحركة (Movement)	الجواب (Answer)	والجهاد (and Jihad)
الزمان (Time)	المقاومة (Resistance)	وبعد (After)	الجهاد (Jihad)
فيما (While)	حول (About)	الموضوع (Subject)	إذ (Then)
الآخرة (Hereafter)	العام (General)	السؤال (Question)	أفغانستان (Afghanistan)
الشيعة (Shiite)	الأستاذ)Profess (or)	لله (God)	أبي (Father of)
العقيدة (Docrine)	العمل (Work)	كان (Was)	العمليات (Operations)
رجل (Man)	العربية (Arabic)	انتهى (Concluded)	العدو (Enemy)
مكة (Mecca)	الإنسان (Human Being)	أعلم (I know)	القتال (Fighting)
السماء (The sky)	الدعوة (Mission)	امرأة (Woman)	التوحيد (Unification)

C. Conclusions and Future Work

We have found that, using stylistic and content features, documents can be automatically identified as belonging to an array of Islamic organizations and ideologies with near-perfect accuracy. The robustness of the results suggests that this methodology can be reliably employed for law enforcement purposes.

REFERENCES

[1] Abbasi, A., and Chen, H., Applying authorship analysis to extremist-group web forum messages, IEEE Intelligent Systems 20(5), pp. 67-75, 2005

[2] Duwairi, R.M. , Machine learning for Arabic text categorization, Journal of the American Society for Information Science and Technology, 57(8), pp.1005-1010, 2006.

[3] Elkourdi, M., Bensaid, A. and Rachidi, T. , Automatic Arabic document categorization based on the Naïve Bayes Algorithm, Proc. of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages, pp. 51-58, 2004.

[4] Forman, G. , An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3(1), pages 1289-1305, 2003.

[5] Genkin, A., Lewis, D. and Madigan, D. , Large-scale Bayesian logistic regression for text categorization, Technometrics 49(3), pp. 291–304, 2007.

[6] Koppel, M., Schler, J. and Argamon, S. , Computational methods in authorship attribution, JASIST 60(1), pp. 9-26, 2009.

[7] Mosteller, F., Wallace, D. L., Inference and Disputed Authorship: The Federalist. Reading, Mass. Addison Wesley, 1964.

[8] Sawaf, H., Zaplo, J. and Ney, H., Statistical classification methods for Arabic news articles, in Proc. of ACL Workshop on Arabic NLP, 2001.

[9] Sebastiani, F. , Machine learning in automated text categorization, ACM Computing Surveys 34(1), pp. 1-47., 2002

[10] Syiam, M., Fayed, Z. and Habib, M., An intelligent system for Arabic text categorization, International Journal of Intelligent Computing and Information Sciences 6(1), pp. 1-19 , 2006.