# Investigations on Standard Arabic Geographical Classification

**Ahmed Abdelali**
New Mexico State University
P.O. Box 30001
Las Cruces, NM 88003
aabdelal@nmsu.edu

**Steve Helmreich**
New Mexico State University
P.O. Box 30001
Las Cruces, NM 88003
shelmreich@psl.nmsu.edu

**Ron Zacharski**
University of Mary Washington
1301 College Avenue
Fredericksburg, VA 22401
raz@umw.edu

## Abstract

This paper reports on a series of studies focused on the geographical classification of Standard Arabic. The aim of these studies was to automatically classify a document based on the author's country of origin. The studies examined documents from newspapers in five countries: Egypt, Libya, Sudan, Syria, and the U.K. using the frequency of common words for classification. We evaluated ten classification algorithms on this task. The best performing algorithms were bagging C4.5, neural network with back propagation, NBTree, and SMO with a polynomial kernel. These methods were over 99% accurate in geographically classifying the documents.

## 1 Introduction[*]

This paper reports on a series of experiments that examine the geographical classification of Modern Standard Arabic. Our goal is to develop automatic methods for determining where a particular Arabic document was written. The methods we have investigated use the frequency of common words to do this classification.

In many applied natural language processing tasks, common words are discarded in a pre-processing step. Often they are considered to constitute *noise* in the system or are considered *non-significant* and are removed. This removal of frequent words is a widespread and standard practice (see, for example, Berry 2004, Dumais and Chen 2000, Gangolly and Yi-Fang 2000, Joachims 1996, Spangler and Kreulen 2008, and van Rijsbergen 1970 among others).

However, there is an equally long history that demonstrates the informativeness of frequent words. One compelling example of this is in the area of stylometrics—the analysis of texts to determine the identity of their authors. In stylometrics the task is to find writer invariant features of text—that is, features that are similar in all the texts of an author but different in the texts of different authors. A number of writer invariants have been identified including syntax, word length, sentence length, vocabulary, and the frequency of function words. Regarding the latter, Mosteller and Wallace in their seminal book on stylometrics (1964), noted that the frequencies of various function words could distinguish the writings of Alexander Hamilton and James Madison. They found, for example, that Hamilton used the word *upon* far more frequently than Madison did—3.24 times per thousand words versus 0.23. 70 function words were used as part of the feature set to classify the documents of the Federalist Papers using a Bayesian approach.

Levison et al. (1968) use the distribution of the particle *de*, the conjunction *kai*, as well as sentence length to argue that the Seventh Letter was not written by Plato, but possibly by Speusippus.[1] Hilton (1990) uses frequent words to determine authorship of the Book of Mormon.

We draw on this idea popular in stylometrics of employing the distribution of common words to

[*] This work has been conducted through participation in CACTUS: Computational Analysis of Cyberterrorism Against the U.S. sponsored by the U.S. Army Test Command, White Sands Missile Range.

[1] This analysis is controversial. See, for example, Deane 1973.

classify documents based on the geographical location of the writer. It has been known for some time that the use of  frequent words such as prepositions varies with regional dialects. (1)-(3) gives examples of these differences. (1a) is an example of the British English use of the preposition *at* in *at the weekend,* where American English speakers would use the preposition *on* as shown in (1b). (2a) illustrates the use of *up* in Black English Venacular (*up my grandmama house*) where speakers of other dialects might use the preposition *to* as shown in (2b). (3a) illustrates the use of *to* by speakers of Outer Bank dialects in contrast to the use of *at* as shown in (3b).

(1) a. *Speaking* ***at*** *the weekend, Cameron said the Torries would also look closely at the National Programme for IT.* [Register 27-4-09]
b. *Speaking* ***on*** *the weekend, Cameron said ...*

(2) a. *See, when I get out of school, I go* ***up*** *my grandmama house.* [Orr 1997]
b. *See, when I get out of school, I go* ***to*** *my grandmama's house.*

(3) a. *She's* ***to*** *the dock.* [Wolfram 2004]
b. *She's* ***at*** *the dock.*

Based on the vast amount of research in this area it is not surprising that someone could identify the region a particular speaker is from based on the distribution and use of common words in their speech. Our research extends this simple notion by seeking the answers to two questions. First, instead of examining specific syntactic constructions, is it possible to identify regions based solely on the count of the different common words? For example, is knowing there were 15 occurrences of *in,* 8 of *up*, and 3 of *to*, sufficient information to identify the author as being from the American Upper Midwest? If it is, it would make it substantially easier to develop computer programs to geographically classify text. Second, instead of identifying dialects (Black English Venacular, Outer Banks dialects,  Appalachian English), is it possible to identify regional differences within one dialect? For example, while speech throughout the U.S. varies, sometimes substantially,  writing appears to have less regional variation than speech.  Given solely the frequency of common words in a written article can we identify the provenance of that article? These two questions (can we classify articles solely on the frequency of common words, and, rather than dialects, can we classify regional variations in one dialect) are what our research seeks to answer.

## 2 Data

We examined these questions using a set of Arabic documents. We collected several corpora for this study.

### 2.1 Newspaper corpus

Our newspaper corpus consisted of 4,167 articles from 5 different countries as shown in table 1.

| Country | Website | # of docs | Avg. doc. Size (kb) |
|---|---|---|---|
| Egypt | ahram.or.eg | 1146 | 21.9 |
| Libya | akhbar-libya.com | 999 | 37 |
| Sudan | almshaheer.com | 749 | 24.3 |
| Syria | thawra.com | 263 | 19.3 |
| UK | asharqalawsat.com | 1010 | 20.3 |

Table 1: Distribution of documents

We represented each document as a vector of the frequencies of common words. We varied the size of the common word list from 58 to 1000 words. The reason for this variation was to determine if classification accuracy would improve with the size of the list.  A subset of these words with their translations is shown below.

| | | | |
|---|---|---|---|
| حول | around | امس | yesterday |
| أي | any | ان | that |
| اعلن | announce | انه | that he |
| حيث | where | او | or |
| الاول | first | اي | any |
| التي | which | ايضا | also |
| الذين | which | قبل | before |
| الى | to | بعد | after |
| اليوم | today | بين | between |
| امام | in front of | حتى | until |

Each of the 4,167 documents of the corpus were converted to this vector format.

### 2.2 Additional newspaper corpus

For the experiment described in §5, we collected a second corpus of newspaper data from three countries: Libya, Sudan and Syria. The distribution of these documents is shown in table 2 below.

| Country | Website | # of docs | Avg. doc. Size (kb) |
|---|---|---|---|
| Libya | ly2day.com | 155 | 4.6 |
| Sudan | rayaam.com | 611 | 7.9 |
| Syria | thisissyria.net | 44 | 3.75 |

Table 2: Distribution of documents of test corpus

The 810 documents' average size was 7.08KB. Unfortunately, as seen in the above table, the file size was not independent of newspaper.

### 2.3 Forum corpus

We also used for testing, a small set of documents we collected from forums in several countries. The distribution of these documents is shown in the following table:

| Country | Website | # of docs | Avg. doc. Size (kb) |
|---|---|---|---|
| Egypt | ahram.org.eg<br>al-ahaly.com<br>almasry-alyoum.com<br>alwafd.org | 50 | 22.9 |
| Libya | alfajraljadeed.com<br>aljamahiria.com<br>alshames.com<br>azzahfalakhder.com | 50 | 12.3 |
| Sudan | p066ezboard.com<br>almshaheer.com<br>midan.com<br>sudaneseonline.com<br>sudanile.com | 48 | 44.6 |
| Syria | fedaa.alwehda.gov.sy<br>furat.alwehda.gov.sy<br>jamahir.alwehda.gov.sy<br>ouruba.alwehda.gov.sy<br>champress.net<br>iqtissadiya.com | 51 | 16.2 |
| UK | alhayat.com<br>asharqalawsat.com | 50 | 29.4 |

Table 3: Distribution of forum documents

## 3 Previous work

In our previous work (Zacharski, et al. 2008) we designed a study to test whether geographical classification of Arabic text is possible using a method based on the distribution of common words. To build a classifier, we trained on the newspaper corpus described in §2.1 using a support vector machine approach.[2] The specific support vector machine algorithm we used was the sequential minimal optimization algorithm (Platt 1998). We evaluated the algorithm using 10-fold cross-validation. We compared the accuracy of using 5 word lists (vector size) differing in how many words they contained: 58, 100, 250, 500, 1000. The 58 word list was a pre-existing one. The remaining lists were constructed by combining the 58 word list with a list of frequent words in the Arabic newspaper corpus.

The results are shown in the following tables. Results range from 92% accurate in classifying documents to over 99%. The rows of the tables represent the actual country the documents were from; the columns represent how the document was classified by our algorithm. For example, in table 4, 1,145 documents from Egypt were correctly classified as being from Egypt; 1 document from Egypt was incorrectly classified as being from Libya. In the next row, 713 of the documents from Sudan were correctly classified as being from Sudan; 1 was incorrectly classified as being from Egypt, 2 from Libya, and 33 from the UK.

| | Egypt | Sudan | Libya | Syria | UK |
|---|---|---|---|---|---|
| Egypt | 1145 | 0 | 1 | 0 | 0 |
| Sudan | 1 | 713 | 2 | 0 | 33 |
| Libya | 21 | 0 | 895 | 0 | 83 |
| Syria | 0 | 0 | 13 | 195 | 55 |
| UK | 1 | 7 | 77 | 30 | 895 |

Table 4: 58 word vector: 92.93% accuracy; $\kappa = 0.899$

[2] In all the studies described in this paper we developed a single classifier to classify all the classes rather than construct separate classifiers for each class.

|  | Egypt | Sudan | Libya | Syria | UK |
|---|---|---|---|---|---|
| Egypt | 1144 | 1 | 0 | 1 | 0 |
| Sudan | 0 | 733 | 0 | 0 | 16 |
| Libya | 4 | 0 | 978 | 0 | 17 |
| Syria | 0 | 0 | 0 | 260 | 3 |
| UK | 0 | 0 | 1 | 7 | 1001 |

Table 5: 100 word vector: 97.41% accuracy; κ = 0.984

|  | Egypt | Sudan | Libya | Syria | UK |
|---|---|---|---|---|---|
| Egypt | 1145 | 0 | 0 | 1 | 0 |
| Sudan | 0 | 746 | 0 | 0 | 3 |
| Libya | 4 | 0 | 989 | 0 | 6 |
| Syria | 0 | 0 | 0 | 252 | 11 |
| UK | 0 | 0 | 3 | 10 | 997 |

Table 6: 250 word vector: 99.09% accuracy; κ = 0.988

|  | Egypt | Sudan | Libya | Syria | UK |
|---|---|---|---|---|---|
| Egypt | 1145 | 0 | 1 | 0 | 0 |
| Sudan | 0 | 748 | 0 | 0 | 1 |
| Libya | 4 | 0 | 992 | 0 | 3 |
| Syria | 0 | 0 | 0 | 260 | 3 |
| UK | 0 | 0 | 1 | 7 | 1001 |

Table 7: 500 word vector: 99.5% accuracy; κ = 0.994

|  | Egypt | Sudan | Libya | Syria | UK |
|---|---|---|---|---|---|
| Egypt | 1145 | 0 | 1 | 0 | 0 |
| Sudan | 0 | 748 | 0 | 0 | 1 |
| Libya | 4 | 0 | 993 | 0 | 2 |
| Syria | 0 | 0 | 0 | 263 | 0 |
| UK | 0 | 0 | 0 | 1 | 1009 |

Table 8: 1k word vector: 99.78% accuracy; κ = 0.997

As the tables show, accuracy improves as the size of the vector increases.

This work suggests that geographically classifying documents based on common words is a promising area to explore. However, it did not address which training methods would lead to accurate classifiers, nor did it adequately examine how well a classifier trained on the newspaper corpus would perform with other genres of written Arabic.

## 4 Comparison of algorithms

Kernel methods, particularly Support Vector Machines, are considered a good approach to problems such as this. One question we might ask is whether other approaches would lead to better performance. Part of our new work focused on comparing the performance of ten algorithms on this dataset. These algorithms are

**Bagging-C4.5.** This method (Breiman 1996) generates ten versions of a C4.5 decision tree classifier (Quinlan 1993) and uses them to produce an aggregate classifier.
**C4.5 decision trees.** This method uses the C4.5 algorithm (J48 implementation) to produce a decision tree classifier (Quinlan 1993).
**Hyperpipes**. This method builds a simple boundary-based classifier (Demiröz and Güenir, 1997; Witten and Frank 1999)
**KNN.** This is the k-nearest neighbor algorithm using three nearest neighbors. The distance is inverse weighted (Aha and Kibler, 1991).
**Naïve Bayes.** This method uses a simple probablistic classifier based on naïve Bayes (John and Langley 1995).
**NBTree**. This method produces a decision tree with naïve Bayes classifiers as leaves (Kohavi 1996).
**NN**. This method assigns the class of the nearest neighbor to the test instance. The distance measure used is Euclidean distance (Aha and Kibler 1991)
**Multilayer Perception.** This method produces a standard neural network classifier. It uses backpropagation (Cybenko, 1989)
**SMO-Poly.** This method uses a support vector machine approach to build a classifier. The specific support vector machine algorithm we used was the sequential minimal optimization algorithm (Platt 1998). A polynomial kernel was used.
**SMO-RBF**. As mentioned immediately above but with an RBF(radial basis function) kernel (Keerthi and Lin 2003).

The accuracy results are shown in table 9.[3]

[3] All experiments conducted with weka 3.5.8 running under Ubuntu 9.04 on a machine with Intel Quad Core Q6600

| | Size of vector | | | | |
|---|---|---|---|---|---|
| | 58 | 100 | 250 | 500 | 1000 |
| Bagging-C4.5 | 96.06 | 97.58 | 99.38 | 99.52 | 99.52 |
| C4.5 | 92.99 | 96.74 | 99.18 | 99.48 | 99.5 |
| Hyperpipes | 72.69 | 84.23 | 94.91 | 97.67 | 97.55 |
| KNN | 94.79 | 97.36 | 98.12 | 97.86 | 97.26 |
| Multilayer P. | 96.21 | 98.06 | 99.18 | 99.52 | 99.62 |
| Naïve Bayes | 79.31 | 91.41 | 96.98 | 98.61 | 98.27 |
| NBTree | 94.31 | 96.85 | 98.8 | 99.25 | 99.4 |
| NN | 95.2 | 97.74 | 98.54 | 98.27 | 97.6 |
| SMO-Poly | 92.23 | 97.41 | 99.09 | 99.5 | 99.78 |
| SMO-RBF | 77.39 | 89.87 | 94.29 | 97.77 | 98.97 |

Table 9: A comparison of classification methods.

By way of comparison, the accuracy statistic in each cell of this table is the same statistic as that reported on in the captions of tables 4-8. These results indicate that Bagging C4.5, C4.5, Multilayer Perceptrons and SMO with a polynomial kernel trained on the 1,000 word vector produce the best results. However, as table 9 shows, if smaller vectors are used, Bagging C4.5 and Multilayer Perceptrons produce better results than SMO.

## 5 Other genres

In our studies up to this point, we have shown that it is possible to classify documents from newspapers with high accuracy. One question we had was whether a classifier trained on this newspaper corpus could classify documents from other genres. To answer this question, we evaluated the classifiers reported on in the previous section on a set of documents from forums. All classifiers were trained on the 1,000 word vector. The details of the collection of forum documents are described in §2.3. The results of this evaluation are shown in table 10. By way of comparison, a classifier trained on the forum posts and evaluated on those posts using 10 fold cross-evaluation was 75.9% accurate.

| | Accuracy | κ |
|---|---|---|
| Bagging-C4.5 | 36.95 | 0.21 |
| C4.5 | 36.14 | 0.2 |
| Hyperpipes | 23.29 | 0.04 |
| KNN | 35.34 | 0.19 |
| Multilayer Perceptron | 19.28 | 0 |
| Naïve Bayes | 27.71 | 0.1 |
| NBTree | 23.29 | 0.04 |
| NN | 34.94 | 0.19 |
| SMO-Poly | 43.78 | 0.3 |
| SMO-RBF | 39.36 | 0.24 |

Table 10: Accuracy of classifiers on forum posts.

As can be seen in the κ values, the classifiers trained on the newspaper corpus have marginal success at classifying forum posts. This may be due to several factors. First, there are likely to be true differences between newspaper documents written in a particular country and forum documents written in that same country. Second, while forums are based in a particular country, the contributors to that forum may be from different countries. This may introduce significant noise that needs to be addressed in future work. Finally, since the forum corpus is small, some variation might be caused by random artifacts.

In our studies up to this point we used documents from one newspaper to represent a country. For example as table 1 shows, all Syrian documents were from thawra.com. It could be argued that the poor performance of the classifiers on the forum corpus was because the classifiers are too narrowly trained on single newspapers thus conflating stylistic differences among newspapers with geographical differences. To test this, we added documents from additional newspapers to the training set and tested these new classifiers with the forum data. This new training data is described in §2.2. The results are shown in Table 11. The results presented in this table suggest, not surprisingly, that adding a different data source improves performance. By way of comparison, just adding 810 documents from the same sources as the original data improves accuracy only about 1.5%. So the near 10% improvement in SMO with the polynomial kernel is a particularly compelling illustration

2.40GHz and 4GB RAM.

of the power of adding data from different sources.

| | Accuracy | κ |
|---|---|---|
| Bagging-C4.5 | 46.18 | 0.33 |
| C4.5 | 51.81 | 0.4 |
| Hyperpipes | 29.71 | 0.12 |
| KNN | 35.34 | 0.19 |
| Multilayer Perceptron | 20.08 | 0 |
| Naïve Bayes | 42.17 | 0.28 |
| NBTree | 26.91 | 0.09 |
| NN | 30.12 | 0.13 |
| SMO-Poly | 53.41 | 0.42 |
| SMO-RBF | 47.39 | 0.34 |

Table 11: Accuracy of new classifiers on forum posts.

## 6 Effects of document size

Our final study examined the effect of document size on classification accuracy. Using SMO with a polynomial kernel and vector sizes of 100 and 1,000, we evaluated classification accuracy on documents of 100 bytes, 500, 1,000, 5,000, and 10,000 derived from the newspaper corpus described in §2.1. The results are shown in table 12.

| Document size | Vector size | |
|---|---|---|
| | 100 | 1k |
| 100 | 59.86 | 90.99 |
| 500 | 90.66 | 99.16 |
| 1k | 93.78 | 99.66 |
| 5k | 98.34 | 99.76 |
| 10k | 97.6 | 99.86 |

Table 12: Accuracy as a function of document size.

As this table shows, the accuracy remains fairly good even for moderately sized documents.

## 7 Conclusion

Our work focused on answering two questions: (1) can we geographically classify documents solely on the frequency of common words, and, (2) rather than dialects, can we classify regional variations in one dialect (for example, can we classify regional differences in Modern Standard Arabic). We developed a series of studies aimed at answering these questions. These studies showed that it is possible to accurately classify newspaper documents solely using the common words in the documents. One study compared the performance of 10 classifiers on this task and provided some evidence that Bagging C4.5, C4.5, and SMO with a polynomial kernel produce the most accurate classifiers. One major limitation of these studies is that they relied on a single data source for each country. Because a single newspaper source was used for each region, it could be argued that the classifiers were classifying the documents based on the newspaper rather than on geographical region. To examine this possibility, we evaluated the performance of the classifier on a different genre: forum posts. The results here are less than compelling; nevertheless the classifier had moderate accuracy on classifying forum posts.[4] We will examine this in more detail in future work using a larger corpus from a wider breadth of sources. Finally, we examined the effect of document size on classification accuracy finding that we could get good classification accuracy even for relatively short documents. These studies suggest that the answer to both questions raised in the beginning sentence of this paragraph is yes: yes we can geographically classify document based on common word frequency and yes we can classify regional differences in Modern Standard Arabic.

This work has direct practical application to intelligence tasks. It may help in determining the author of an anonymous document. For example, a geographical classifier can be used as one module of a system designed to detect cyber terrorist threats against the U.S. by aiding in the identification of the source of the threat. Finally, many Arabic scholars (Shukri B. Abed, p.c.) believe there are no regional variations of Modern Standard Arabic. The work reported on here provides some support for the alternative view that there are regional variations (see, for example, Ibrahim and Ibrahim, 2009 and Abdelali, 2004). Future work using larger corpora from a broad number of sources may provide stronger evidence for this position.

[4] The best classifier on this task had a κ of 0.42 which is generally considered 'moderate' agreement (Viera and Garrett 2005).

## References

Abdelali, Ahmed. 2004. Localization in Modern Standard Arabic. *Journal of the American Society for Information Science and Technology* 55(1):23-8.

Aha, D. and D. Kibler. 1991. Instance-based learning algorithms. *Machine learning* 6.37-66.

Berry, Michael W. 2004. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer, Berlin.

Breiman, Leo. 1996. Bagging Predictors. *Machine Learning* 24(2).123-140.

Cybenko, G. 1989. Approximation by superpostions of a sigmoidal function. Mathematics of Control, Signals, and Systems. 2(4).303-314.

Deane, Philip. 1973. Stylometrics do not Exclude the Seventh Letter. *Mind* 82.325: 113-117.

Demiröz, Gülson, and H. Altay Güenir. 1997. Classification by Voting Feature Intervals. Proceedings of the European Conference on Machine Learning. 85-92.

Dumais, Susan, and Hao Chen. 2000. Hierarchical classification of web content. *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* edited by Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, 256-263. ACM Press, New York.

Friedman, Jerome H.; Jon Luis Bentley; and Raphael Ari Finkel. 1977. An Algorithm for Finding Best Matches in Logarithmic Expected Time. ACM *Transactions on Mathematics Software* 3(3).209-226.

Gangolly, Jagdish, and Wu Yi-Fang. 2000. On the Automatic Classification of Accounting concepts: Preliminary Results of the Statistical Analysis of Term-Document Frequencies. *New Review of Applied Exert Systems and Emerging Technologies*, 6.81-88.

Garner, S. R. 1995. WEKA: The Waikato Environment for Knowledge Analysis. Proceedings of the New Zealand Computer Science Research Students Conference. 57-64.

George, John, and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In: *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 338-345.

Gundel, Jeanette; Nancy Hedberg; and Ron Zacharski. 2000. Statut cognitif et forme des anaphoriques indirects. *Verbum* 22.79-102.

Hilton, John L. 1990. On verifying wordprint studies: Book of Mormon authorship. *Book of Mormon Authorship Revisited: The evidence for ancient origins*, edited by Noel Reynolds, 225-253. Foundation for Ancient Research and Mormon Studies, Provo Utah.

Ibrahim, Zeinab and Zaynab Ibrahim. 2009. Beyond Lexical Variation in Modern Standard Ararbic: Egypt, Lebanon, and Morocco. Cambridge Scholars Publishing, New Castle upon Tyne.

Joachims, T. 1996. A probabilistic analysis of the Rocchio Algorithm with TFIDF for text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, 143-151.

Keerthi, S. S. and C. J. Lin. 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*. 15(7).1667-1689.

Kohavi, Ron. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision tree hybrid. *Second International Conference on Knowledge Discovery and Data Mining*. 202-207.

Levison, M., A. Q. Morton, and A. D. Winspear. 1968. The Seventh Letter of Plato. *Mind* 77.309-325.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2.159-165.

Moore, Andrew. 1991. A tutorial on kd-trees. University of Cambridge Computer Laboratory Technical Report No. 209.

Mosteller, Frederick, and David K. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley Publishing, Reading, MA.

Orr, Eleanor. 1997. *Twice as less: Black English and the performance of black students in mathematics and science.* Norton, New York.

Platt, John C. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning* edited by Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, 185-208. MIT Press, Cambridge.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA.

Spangler, Scott, and Jeffrey Kreulen. 2008. *Mining the Talk: Unlocking the Business Value in Unstructured Information*. IBM Press, Upper Saddle River, New Jersey.

Tong, Simon; Uri Lerner; Amit Singhal; Paul Haahr; and Stephen Baker. 2008. Locating meaningful stopwords or stop-phrases in keyword-based retrieval systems. United States Patent http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,409,383.PN.&OS=pn/7,409,383&RS=PN/7,409,383

van Rijsbergen, C.J. 1979. *Information Retrieval*. Butterworths, London

Viera, Anthony, and Joanne M. Garrett. 2005. Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine 37(5).360-363.

Witten, Ian H. and Eibe Frank. 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman,

Wolfram, Walt. 2004. Rural and ethnic varieties in the Southeast: morphology and Syntax. *A Handbook of varieties of English.* ed. By Bernd Kortmann, Kate Burridge, Rajend Mesthrie, Edgar Schneider, and Clive Upton. de Gruyter. New York. 281-302.

Zacharski, Ron; Ahmed Abdelali; Steve Helmreich; and Jim Cowie. 2008. Linguistic Dumpster Diving: Geographical Classification of Arabic Text. Proceedings of the Chicago Colloquia on Digital Humanities and Computer Science.

## Appendix: Details of classifiers

Below we list for each classifier the Weka (Garner 1995) command and parameters that were used.

**Bagging C4.5**:

```
weka.classifiers.meta.Bagging -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
```

**C4.5**:

```
weka.classifiers.trees.J48 -C 0.25 -M 2
```

**Hyperpipes**:

```
HyperPipes
```

**KNN**:

```
weka.classifiers.lazy.IBk -K 3 -W 0 -I -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
```

**Multilayer Perceptron:**

```
weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
```

**naïve Bayes**:

```
weka.classifiers.bayes.NaiveBayes
```

**NBTree**:

```
NBTree
```

**NN**:

```
IB1
```

**SMO-Poly**:

```
weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
```

**SMO-RBF**:

```
weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01"
```