New Methods for Attribution of Rabbinic Literature

Moshe Koppel   Dror Mughaz   Navot Akiva
{koppel,myghaz,navot}@cs.biu.ac.il
Dept. of Computer Science
Bar-Ilan University

Introduction

In this paper, we will demonstrate how recent developments in the nascent field of automated text categorization can be applied to Hebrew and Hebrew-Aramaic texts. In particular, we illustrate the use of new computational methods to address a number of scholarly problems concerning the classification of rabbinic manuscripts. These problems include ascertaining answers to the following questions

1. Which of a set of known authors is the most likely author of a given document of unknown provenance?
2. Were two given corpora written/edited by the same author or not?
3. Which of a set of documents preceded which and did some influence others?
4. From which version (manuscript) of a document is a given fragment taken?

We will apply our techniques to a number of representative problems involving corpora of rabbinic texts.

Text Categorization

Text categorization is one of the major problems of the field of machine learning (Sebastiani 2002). The idea is that we are given two or more classes of documents and we need to find some formula (usually called a "model") that reflects statistical differences between the classes and that can then be used to classify a new document. 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 man or a woman, as having been written by one of a given set of candidate authors and so forth.

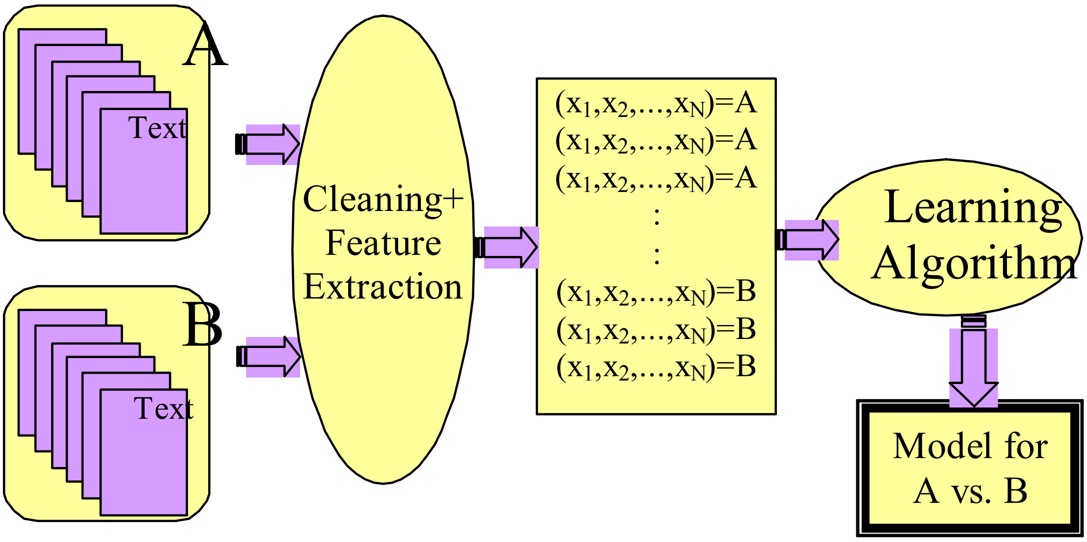

Figure 1: Architecture of a text categorization system.

In Figure 1 we show the basic architecture of a text categorization system in which we are given examples of two classes of documents, Class A and Class B. The first step, document representation, involves defining a set of text features which might potentially be useful for categorizing texts in a given corpus (for example, words that are neither too common nor too rare) 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. Optionally, one may then use various criteria for reducing the dimension of the feature vectors (Yang & Pedersen 1997).

Once documents have been represented as vectors, there are a number of learning algorithms that can be used to construct models that distinguish between vectors representing documents in Class A and vectors representing documents in Class B. Yang (1999) compares and assesses some of the most promising algorithms, which include k-nearest-neighbor, neural nets, Winnow, SVM, etc. One particular type of model which is easy to understand, and which we use in this paper, is known as a linear separator. A linear separator works as follows: we assign to each feature of a text a certain number of points to either Class A or to Class B. (The class to which points are assigned and the precise number of points assigned are determined by the learning algorithm based on the training documents.) Then a new document is classified by scanning it and counting how many points it contains for each class. The class with most points in the document is the class to which the document is assigned.

Style-Based Text Categorization

Driven largely by the problem of Internet search, the text categorization literature has dealt primarily with classification of texts by topic: to which category in some directory of topics should a document (typically, a web page) be assigned. There has, however, been a considerable amount of research on authorship attribution, which is what concerns us in this paper. Most of this work has taken place within what is often called the "stylometric" community (Holmes 1998, McEnery & Oakes 2000), which has tended to use statistical methods substantially different in flavor from those typically used by researchers in the machine learning community. Nevertheless, in recent years machine learning techniques have been used with increasing frequency for solving style-based problems. The granddaddy of such works is that of Mosteller and Wallace who applied Naïve Bayes to solve the problem of the Federalist Papers. Other such works include that of Matthews and Merriam (1993) on the works of Shakespeare, Argamon et al (1998) on news stories, Koppel et al (2002) on gender, Wolters and Kirsten on genre (2001), deVel (2001) on email authorship, Stamatatos et al (2001) on Greek.

Classification according to topic is a significantly easier problem than classifying according to author style. The kinds of features which researchers use for categorizing according to topic typically are frequencies of content 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 author style one needs to use precisely those linguistic features that are content-independent. We wish to find those stylistic features of a given author's writing that are independent of any particular topic. Thus, in the past researchers have used for this purpose lexical features such as function words (Mosteller & Wallace 1964), syntactic features (Baayen et al 1996, Argamon et al 1999, Stamatatos et al 2001), or complexity-based feature such as word and sentence length (Yule 1938). As we will see, different applications call for different types of features.

Hebrew texts present special problems in terms of feature selection for style-based classification. In particular, function words tend to be conflated into word affixes in Hebrew, thus decreasing the number of function words but increasing the amount of morphological features that can be exploited. The richness of Hebrew morphology also renders part-of-speech tagging a much messier task in Hebrew than in other languages, such as English, in which each part-of-speech typically is represented as a separate word. In any case, we did not use a Hebrew part-of-speech tagger for this study. A good deal of work has been done by Radai (1978, 1979, 1982) on categorization of Biblical documents but Radai's work was not done in the machine learning paradigm used in this paper.

Our Approach

In this paper we will solve four problems all involving texts in Hebrew-Aramaic.

Problem 1: We are given responsa (letters written in response to legal questions) of two authorities in Jewish law, Rashba and Ritba. Both lived in Spain in the thirteenth century, where Ritba was a student of Rashba. Their styles are regarded as very similar to each other. The object is to identify a given responsum as having been authored by Rashba or Ritba.

Problem 2: We are given one corpus written by a nineteenth century Baghdadi scholar, Ben Ish Chai, and another corpus believed to have been written by him under a pseudonym. We need to determine if the same person wrote the two corpora or not.

Problem 3: We are given three sub-corpora of the classic work of Jewish mysticism, Zohar. Scholars are uncertain whether a single author authored the three corpora and, if not, which corpora influenced which others. We will resolve the authorship issue and propose the likeliest relationship between the corpora.

Problem 4: We are given four manuscripts, one printed version and three hand-written by different scribes, of the same tractate of the Babylonian Talmud. The object is to determine from which manuscript a given fragment is taken. (We are given the text of the fragment, not the original so that handwriting is not relevant.)

Problem 1: Authorship Attribution

The problem of authorship attribution is the simplest one we will consider in this paper and its solution forms the basis for the solutions of all the other problems. It is a straightforward application of the techniques described above: we are given the writings of a set of authors and are asked to classify previously unseen documents as belonging to one or the other of these authors.

To illustrate how this is done, we consider the problem of determining whether a given responsum was written by Rashba, a leading thirteenth century rabbinic scholar, or by his student, Ritba. We consider this problem merely as an exercise; to the best of our knowledge, there are no extant responsa of disputed authorship in which these two scholars are the candidate authors. We are given 209 responsa from each Ritba and Rashba. We select a list of lexical features as follows: the 500 most frequent words in the corpus are selected and all those that are deemed content-words are eliminated manually. We are left with 304 features. Strictly speaking, these are not all function words but rather words that are typical of the legal genre generally without being correlated with any particular sub-genre. Thus a word like שאלה would be allowed, although in other contexts it would not be considered a function word.

An important aspect of this experiment is the pre-processing that must be applied to the text before vectors can be constructed. Since the texts we have of the response have undergone editing, we must make sure to ignore possible effects of differences in the texts resulting from variant editing practices. Thus, we expand all abbreviations and unify variant spellings of the same word.

After representing each of our training examples as a numerical vector, we use as our learning algorithm a generalization of the Balanced Winnow algorithm of Littlestone (1987) that has previously been shown to be effective for text-categorization by topic (Lewis et al 1996, Dagan et al 1997) and by style (Koppel et al 2003).

In order to test the accuracy of our methods, we need to test the models on documents that were not seen by the computer during the training phase. To do this properly, we use a technique known as five-fold cross-validation, which works as follows: We take all the documents in our corpus and randomly divide them into five sets. We then choose four of the sets and learn a model that distinguishes between Rashba and Ritba. Once we have done this we take the fifth set and apply the learned model to this set to see how well the model works. We do this five times, each time holding out a different one of the five sets as a test set. Then we record the accuracy of our models overall at classifying the test examples.

Application of the Balanced Winnow algorithm on our full feature set in five-fold cross-validation experiments yielded test accuracy of 85.8%. After removing features which received low weights and then re-running Balanced Winnow from scratch, we obtained accuracy of 90.5%**.**

It is interesting to note the features that turn out to be most effective for distinguishing between these authors. The word ולפיכך is used over 30 times more frequently by Rashba, while הנזכר is used over 40 times more frequently by the Ritba. Similarly, אמרת and שאלת are used significantly more by Rashba. Table 1 shows a number of features that are used with significantly different frequency by Rashba and Ritba, respectively. Note that Rashba tends to employ more second person and plural first person pronouns than does Ritba. This might be taken as evidence of attempts by Rashba to encourage "involvedness" (Biber et al 1998) on the part of his respondents.

Ritba	Rashba	feature
5.18	0.86	פשוט
7.59	0.96	שאלה
45.14	0.96	הנזכר
6.65	1.50	דעתי
2.50	0.64	מפורש
10.18	3.63	דעת
5.35	13.68	שאלת
0.78	4.60	דגרסינן
0.52	3.74	והיינו
1.04	3.74	אמרת
0.17	5.45	ולפיכך
1.29	2.89	שאנו

Table 1: Frequencies (per 10000 words) of various words in the Rashba and Ritba corpora, respectively.

We have run other similar experiments (Mughaz 2003) too numerous to present in detail here. For example, we have found that glosses of the Tosafists can be classified according to their provenance (Evreaux, Sens, Germany) with accuracy of 90% (see

Urbach 1954). Likewise, sections of Midrash Halakhah can be classified as originating in the school of Rabbi Aqiba or the school of Rabbi Yishmael with accuracy of 95% (see Epstein 1957).

Problem 2: Unmasking Pseudonymous Authors

The second problem we consider is that of authorship verification. In the authorship verification problem, we are given examples of the writing of a single author and are asked to determine if given texts were or were not written by this author. As a categorization problem, verification is significantly more difficult than attribution and little, if any, work has been performed on it in the learning community. As we have seen, when we wish to determine if a text was written by, for example, Rashba or Ritba, it is sufficient to use their respective known writings, to construct a model distinguishing them, and to test the unknown text against the model. If, on the other hand, we need to determine if a text was written by Rashba or not, it is very difficult – if not impossible – to assemble an exhaustive, or even representative, sample of not-Rashba. The situation in which we suspect that a given author may have written some text but do not have an exhaustive list of alternative candidates is a common one.

The particular authorship verification problem we will consider here is a genuine literary conundrum. We are given two nineteenth century collections of Hebrew-Aramaic responsa. The first, *RP (Rav Pe'alim)* includes 509 documents authored by an Iraqi rabbinic scholar known as Ben Ish Chai. The second, *TL (Torah Lishmah)* includes 524 documents that Ben Ish Chai, claims to have found in an archive. There is ample historical reason to believe that he in fact authored the manuscript but did not wish to take credit for it for personal reasons (Ben-David 2003). What do the texts tell us?

The first thing we do is to find four more collections of responsa written by four other authors working in roughly the same area during (very) roughly the same period. These texts are *Zivhei Zedeq* (Iraq, nineteenth century), *Shoel veNishal* (Tunisia, nineteenth century), *Darhei Noam* (Egypt, seventeenth century), and *Ginat Veradim* (Egypt, seventeenth century). We begin by checking whether we able to distinguish one collection from another using standard text categorization techniques. We select a list of lexical features as follows: the 200 most frequent words in the corpus are selected and all those that are deemed content-words are eliminated manually. We are left with 130 features. After pre-processing the text as in the previous experiment, we constructed vectors of length 130 in which each element represented the relative frequency (normalized by document length) of each feature.

We then used Balanced Winnow as our learner to distinguish pairwise between the various collections. Five-fold cross-validation experiments yield accuracy of greater than 95% for each pair. In particular, we are able to distinguish between *RP* and *TL* with accuracy of 98.5%.

One might thus be led to conclude that *RP* and *TL* are by different authors. It is still possible, however, that in fact only a small number of features are doing all the work of distinguishing between them. The situation in which an author will use a small number of features in a consistently different way between works is typical. These differences might result from thematic differences between the works, from differences in genre or purpose, from chronological stylistic drift, or from deliberate attempts by the author to mask his or her identity.

In order to test whether the differences found between *RP* and *TL* reflect relatively shallow differences that can be expected between two works of the same author or reflect deeper differences that can be expected between two different authors, we invented a new technique that we call *unmasking* (Koppel et al 2004, Koppel and Schler 2004) that works as follows:

We begin by learning models to distinguish *TL* from each of the other authors including *RP*. As noted, such models are quite effective. In each case, we then eliminate the five highest-weighted features and learn a new model. We iterate this procedure ten times. The depth of difference between a given pair can then be gauged by the rate with which results degrade as good features are eliminated.

The results (shown in Figure 1) could not be more glaring. For *TL* versus each author other than *RP*, we are able to distinguish with gradually degrading effectiveness as the best features are dropped. But for *TL* versus *RP*, the effectiveness of the models drops right off a shelf. This indicates that just a few features, possibly deliberately inserted as a ruse or possibly a function of slightly differing purposes assigned to the works, distinguish between the works. For example, the frequency (per 10000 words) of the word זה in *RP* is 80 and in *TL* is 116. A cursory glance at the texts is enough to establish why this is the case: the author of *TL* ended every responsum with the phrase והיה זה שלום, thus artificially inflating the frequency of these words. Indeed the presence or absence of this phrase alone is enough to allow highly accurate classification of a given responsum as either *RP* or *TL*. Once features of this sort are eliminated, the works become indistinguishable – a phenomenon which does not occur when we compare *TL* to each of the other collections. In other words, many features can be used to distinguish *TL* from works in our corpus other than *RP*, but only a few distinguish *TL* from *RP*. Most features distribute similarly in *RP* and *TL*. A wonderful illustrative example of this is the word וכו', the respective frequencies of which in the various corpora are as follows: *TL:29 RP:28 SV:4 GV:4 DN:41 ZZ:77*

We have shown elsewhere (Koppel and Schler 2004), that the evidence offered in Figure 1 is sufficient to conclude that the author of RP and TL are one and the same: Ben Ish Chai.

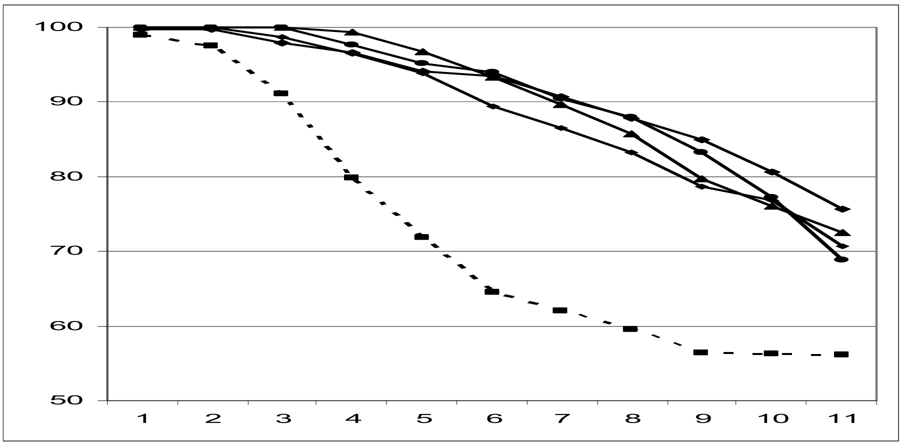

Fig. 1: Accuracy (*y*-axis) on training data of learned models comparing *TL* to other collections as best features are eliminated, five per iteration (*x*-axis). Dotted line on bottom is *RP* vs. *TL*.

Problem 3: Chronology and Dependence

Given three or more corpora, we can attempt to learn dependencies among the corpora by checking pairwise similarities. To illustrate what we mean, we consider three sub-corpora of Zohar, the central text of Jewish mysticism:

HaIdra (47 passages from Zohar vol. 3, pp. 127b-141a; 287b-296b)
Midrash HaNe'elam (67 passages from Zohar vol. 1, pp. 97a-140a)
Raya Mehemna (100 passages from Zohar vol. 3, 215b-283a)

For simplicity, we will refer to these three corpora as I, M and R, respectively. These sub-corpora are of disputed provenance and their chronology and cross-influence are not well established.

Lexical features were chosen in a similar fashion to that described above. Separate models were constructed to distinguish between each pair from among the three corpora. In five-fold cross-validation experiments on each pair, unseen documents were classified with approximately 98% accuracy. In addition, degradation using unmasking is slow. From this we conclude that these corpora were written by three different authors.

The next stage of the experiment is an attempt to determine the relationship between the three corpora. We learn models to distinguish two of the corpora from each other and then use this model to classify the third corpora as more similar to one or the other. In this way we hope to determine possible dependencies among the corpora.

In our initial experiments, absolutely nothing could be concluded because in each of the three experiments the passages of the third corpus seemed to split about evenly between being more similar to the first as to the second. We then ran the experiment again but this time using grammatical prefixes and suffixes as features. Using the expanded feature set, we were able to pair-wise distinguish between the corpora with the same 98% accuracy as with the original lexical feature set. However, the results of the second experiment changed dramatically. When we learn models distinguishing between R and M and then use them to classify I, all I passages are classified as closer to R. Similarly, when we learn models distinguishing between R and I and then use them to classify M, all M passages are classified as closer to R. However, when we learn models distinguishing between M and I and then use them to classify R, the results are ambiguous.

The reason for this is rooted in the fact that, like our other corpora, Zohar is written in a dialect that combines Aramaic and Hebrew. One of the main distinguishing features of Hebrew versus Aramaic is the use of certain affixes. For example, in Hebrew the plural noun suffixes are ות and ים, while in Aramaic ין and נא are used. Similarly, in Hebrew the word *which* is incorporated as the prefix ש while in Aramaic ד is used. We find (see Table 2) that M is characterized by a large number of Hebrew affixes and I is distinguished by a large number of Aramaic affixes. R falls neatly in the middle.

M	R	I	feature
8.12	0.17	0.00	שה*
2.92	4.24	3.87	וש*
4.37	11.61	6.44	*ית
7.36	25.73	13.70	*הו
3.75	4.41	8.81	*וי
4.37	11.61	10.38	וב*
15.82	11.54	6.21	*ות
9.44	6.53	3.02	*את
33.66	21.42	9.57	*ים
7.84	14.87	32.92	*נא
10.73	4.65	2.34	וי*
17.63	61.20	87.28	*ין
9.59	2.34	0.72	של*

Table 2: Frequencies of prefixes and suffixes in I/M/R)

A number of possible conclusions might be drawn from this. For example, the phenomena uncovered here might support the hypothesis that R lies chronologically between M and I. However, scholars of this material believe that a more likely interpretation is that M and I were cotemporaneous and independent of each other and that R was subsequent to both and may have drawn from each of them.

Problem 4: Assigning manuscript fragments

Our final experiment is a version of an attribution experiment. However, in this case we wish to distinguish between different versions of the same text. The question is whether we can exploit differences in orthographic style to correctly assign some text fragment to the manuscript from which it was taken.

In our experiment, we are given four versions of the same Talmudic text (tractate Rosh Hashana of the Babylonian Talmud), each version having been transcribed by a different scribe. We break each of the four manuscripts into 67 fragments (corresponding to pages in the printed version). The object is to determine from which version a given fragment might have come.

Note that since we are distinguishing between different versions of the same texts, we can't realistically expect lexical or morphological features to distinguish very well. After all, the texts consist of the same words. Rather, the features that are likely to help here are precisely those that were disqualified in our earlier experiments, namely, orthographic ones.

Rather than identify these features manually, we proceeded as follows. First, we simply gathered a list of all lexical features that appeared at least ten times in the texts. Variant spellings of the same word were treated as separate features. In order to identify promising features, we used an "instability" measure (Koppel et al, 2003) that grants a high score to a feature that appears with different frequency in different versions of the same document.

Specifically, let $\{d_1,d_2,\ldots,d_n\}$ be a set of texts (in our case n=67) and let $\{d_i^1,d_i^2,\ldots,d_i^m\}$ be m > 1 different versions of d_i (in our case m=4). For each feature c, let c_i^j be the relative frequency of c in document d_i^j. For multiple versions of a single text d_i, let $k_i = \Sigma_j\, c_i^j$ and let $H(c_i) = -\Sigma_j\, [(c_i^j/\, k_i)\log\, (c_i^j/\, k_i)]]/\log m$. (We can think of $c_i^j/\, k_i$ as the probability that a random appearance of c in d_i is in version d_i^j so that $H(c_i)$ is just the usual entropy measure.) Thus, for example, if a feature c assumed the identical value in every version of a document d_i, $H(c_i)$ would be 1. To extend the definition to the whole set $\{d_1,d_2,\ldots,d_n\}$, let $K = \Sigma_i\, k_i$ and let $H(c) = \Sigma_i\, [(k_i/K) * H(c_i)]$. Finally, let

H'(*c*) = 1-H(*c*). H'(*c*) does exactly what we want: features the frequency of which varies in different versions of the same document score higher than those that have the same frequency in each version.

We then ranked all features according to H'(*c*). Those that ranked highest were those that permitted variant orthographic representations. In particular, some scribes used abbreviations or acronyms or non-standard spellings in places where other scribes did not. We choose as our feature set the 200 highest-ranked features according to H' in the training corpus. Using Naïve Bayes on this feature set in five-fold cross-validation experiments yielded accuracy of 85.4%.

Thus, by and large, we are able to correctly assign a fragment with its manuscript of origin. This work recapitulates and extends in automated fashion, a significant amount of research carried out manually by scholars of Talmudic literature (Friedman 1996). Among the main distinguishing features we find different substitutions for the Name of God (ה', יי', יי"י, '''), variant abbreviations (דכת', דכתי', דכתיב), and a number of acronyms (א"ר, ת"ר, מ"ט, ת"ש, ס"ד) used in some manuscripts but not in others.

There is one major limitation to the approach we used here. We assume that within a given manuscript the frequency of a given feature is reasonably invariant from fragment to fragment. This is only true if we are considering various versions of a single thematically homogeneous text. If we wish to train on versions of various texts as a basis for identifying the scribe/editor of a manuscript of a different text, we need make a more realistic assumption. This can be done by normalizing our feature frequencies differently: we must count the number of appearances of a particular orthographic variant of a word in a manuscript fragment relative to the total number of appearances of all variants of that word in the fragment. This value should indeed remain reasonably constant for a single scribe/editor across all texts.

Conclusions

We have shown that the range of issues considered in the field of text categorization can be significantly broadened to include problems of great importance to scholars in the humanities. Methods already used in text categorization require a bit of adaptation to handle these problems. First, the proper choice of feature sets (lexical, morphological and orthographic) is required. In addition, juxtaposition of a variety of classification experiments can be used to handle issues of pseudonymous writing, chronology and other problems in surprising ways. We have seen that for a variety of textual problems concerning Hebrew-Aramaic texts, proper selection of feature sets combined with these new techniques can yield results of great use to scholars in these areas.

References

Argamon-Engelson, S., M. Koppel, G. Avneri (1998). Style-based text categorization: What newspaper am I reading?, in Proc. of AAAI Workshop on Learning for Text Categorization, 1998, pp. 1-4

Baayen, H., H. van Halteren, F. Tweedie (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing, 11, 1996.

Ben-David, Y. L. (2002). *Shevet mi-Yehudah* (in Hebrew), Jerusalem, 2002 (no publisher listed)

Biber, D., S. Conrad, R. Reppen (1998). *Corpus Linguistics: Investigating Language Structure and Use*, (Cambridge University Press, Cambridge, 1998).

Dagan, I., Y. Karov, D. Roth (1997), Mistake-driven learning in text categorization, in EMNLP-97: 2nd Conf. on Empirical Methods in Natural Language Processing, 1997, pp. 55-63.

de Vel, O., A. Anderson, M. Corney and George M. Mohay (2001). Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64

Epstein, Y.N. (1957). *Mevo'ot le-Sifrut ha-Tana'im*, Jerusalem, 1957

Friedman, S. (1996) The Manuscripts of the Babylonian Talmud: A Typology Based Upon Orthographic and Linguistic Features. In: Bar-Asher, M. (ed.) *Studies in Hebrew and Jewish Languages Presented to Shelomo Morag* [in Hebrew], p. 163-190. Jerusalem, 1996.

Holmes, D. (1998). The evolution of stylometry in humanities scholarship, Literary and Linguistic Computing, 13, 3, 1998, pp. 111-117.

Koppel, M., N. Akiva and I. Dagan (2003), A corpus-independent feature set for style-based text categorization, in Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico.

Koppel, M., S. Argamon, A. Shimony (2002). Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17,4, Nov. 2002, pp. 401-412

Koppel, M., D. Mughaz and J. Schler (2004). Text categorization for authorship verification in Proc. 8th Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, 2004.

Koppel and J. Schler (2004), Authorship Verification as a One-Class Classification Problem, to appear in Proc. of ICML 2004, Banff, Canada

Lewis, D., R. Schapire, J. Callan, R. Papka (1996). Training algorithms for text classifiers, in Proc. 19th ACM/SIGIR Conf. on R&D in IR, 1996, pp. 306-298.

Littlestone, N. (1987). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, Machine Learning, 2, 4, 1987, pp. 285-318.

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry : An application to the works of Shakespeare and Fletcher. Literary and Linguistic computing, 8(4):203-209.

McEnery, A., M. Oakes (2000). Authorship studies/textual statistics, in R. Dale, H. Moisl, H. Somers eds., Handbook of Natural Language Processing (Marcel Dekker, 2000).

Merriam, T. and Matthews, R. (1994). Neural computation in stylometry : An application to the works of Shakespeare and Marlowe. Literary and Linguistic computing, 9(1):1-6.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist.* Reading, Mass. : Addison Wesley, 1964.

Mughaz, D. (2003). *Classification Of Hebrew Texts according to Style*, M.Sc. Thesis, Bar-Ilan University, Ramat-Gan, Israel, 2003.

Radai, Y. (1978). *Hamikra haMemuchshav: Hesegim Bikoret uMishalot* (in Hebrew), Balshanut Ivrit 13: 92-99

Radai, Y. (1979). *Od al Hamikra haMemuchshav* (in Hebrew), Balshanut Ivrit 15: 58-59

Radai, Y. (1982). *Mikra uMachshev: Divrei Idkun* (in Hebrew), Balshanut Ivrit 19: 47-52

Sebastiani, F. (2002). Machine learning in automated text categorization, ACM Computing Surveys 34 (1), pp. 1-45

Stamatatos, E., N. Fakotakis & G. Kokkinakis, (2001). Computer-based authorship attribution without lexical measures, Computers and the Humanities 35, pp. 193—214.

Tishbi, Y. (1949). *Mishnat haZohar* (in Hebrew), Magnes: Jerusalem, 1949.

Urbach, E. E. (1954). *Baalei haTosafot* (in Hebrew), Bialik: Jerusalem, 1954.

Wolters, M. and Kirsten, M. (1999): Exploring the Use of Linguistic Features in Domain and Genre Classification, Proceedings of the Meeting of the European Chapter of the Association for Computational Linguistics

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, Vol 1, No. 1/2, pp 67--88, 1999.

Yang, Y. and Pedersen, J.O. (1997). A comparative study on feature selection in text categorization, Proceedings of ICML-97, 14th International Conference on Machine Learning,412--420

Yule, G.U. (1938). "On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship", Biometrika, 30, 363-390, 1938.