2011 10th International Conference on Machine Learning and Applications

Machine Learning for Author Affiliation within Web Forums

Using Statistical Techniques on NLP Features for Online Group Identification

Jeffrey Ellen[*], Shibin Parameswaran[*]
Space and Naval Warfare Systems Center Pacific
United States Navy
San Diego, CA USA
jeffrey.ellen@navy.mil, shibin.parameswaran@navy.mil

Abstract— **Although there have been previous studies performing authorship attribution to a specific individual; we find a shortage of efforts to group authors based on their affiliations. This paper presents our work on classification of website forum posts by the author's group affiliation. Specifically, we seek to classify translated website forum posts by the (inferred) political affiliation of the author. The two datasets that we attempt to classify consist of real-world data discussing current issues -- Israeli/Palestinian dialogue (BitterLemons corpus) and translated Extremist/Moderate forum entries (from internet websites). To achieve our goal of reliable authorship affiliation, we extract term frequency-based features (that are conventional in document classification) along with less commonly used linguistic style-based features. The resulting set of stylometric features are then utilized in two widely used supervised classification algorithms, namely k-Nearest Neighbor algorithm and Support Vector Machines. Specifically, we used k-NN with cosine distance and Support Vector Machines with two different kernel functions. In addition to the popular RBF kernels, we also evaluate the applicability and performance of the recently introduced arc-cosine kernels for group affiliation. The results of our experiments show strong performance across a range of pertinent metrics.**

Keywords: Text Classification; Stylometrics; Natural language processing; feature extraction; feature combination; Support vector machines; k-nearest neighbor; arccosine kernels

I. INTRODUCTION

In this paper, we work towards classifying website forum posts based on their authorship affiliation to a particular group. We experimented with the most popular and proficient machine learning methodologies currently available for text classification, and contrasted results against some of the latest available techniques. We feel our findings are of interest to the machine learning research community because:

- There is a lack of current published research regarding training algorithms to identify group affiliation
- We experimented with less commonly used features and found that they consistently boosted performance for this task
- We were able to obtain extremely high performance on our test data sets.

A. Background and Motivation

Here we describe algorithms that we have trained to ascertain group membership or affiliation of the author of a website forum post.

There are many domains where knowing the affiliation of an anonymous/pseudonymous author would be of interest. Our work would be valuable to market researchers, who would better be able to ascertain groupwise demographic information, especially in combination with sentiment analysis. For example, if authors reviewing a specific movie or product could be grouped, this could result in more effective advertising, or even refining a product defect, or expanding a successful product. (Are all the negative reviews of this product from mothers? Can we fix the perceived safety/friendliness of this product?) The inverse of this could be applied within the financial/investment sector to help ferret out shills. (Was this glowing review of a stock written by a marketer or insider?) Within the educational field, this type of analysis could be used to help determine authenticity of students' submissions. (Was this prose likely authored by a student or a non-student?) Individuals involved with the legal/discovery process might be interested in impartial, algorithmic analysis of text with respect to group affiliation, along the lines of the insider/shill financial example.

We are concerned specifically with assisting intelligence analysts from a US Department of Defense perspective. In our case, we are seeking to perform group affiliation for purposes such as locating and tracking whether posts of an inflammatory nature within a certain community are being authored by actual members of the community, or by infiltrators and propagandists trying to sway a stable population or political group. Similar to the market research example earlier, in combination with sentiment analysis, improved groupwise demographics might help better guide foreign policy. The work in this paper is an important first step towards achieving some of these goals, more broadly referred to as "Cognitive Information Operations" [1].

B. Related Work

Much current work on document classification focuses on sorting documents by subject, which naturally has a heavy reliance on term frequency (TF) or a variant thereof. We

[*] Equally contributing authors

Work sponsored by the Office of Naval Research Code 30

978-0-7695-4607-0 2011
U.S. Government Work Not Protected by U.S. Copyright
DOI 10.1109/ICMLA.2011.90

IEEE computer society

introduce additional Natural Language Processing (NLP) features based on our research into authorship attribution as well as additional features based on our own intuition.

Authorship attribution [25], ‘fingerprinting’ text or ‘writeprinting’ [21] is a related research field which has numerous ongoing investigations, most of which are based on attempts to construct a ‘signature’ to identify a particular author [3][8]. There is considerable progress in this area, including a study with success identifying an individual forum post author out of a group of 20 peers [6]. Our research is a more generalized case of authorship attribution; we are trying to determine identifying characteristics for affiliation at a group level. Please note that we are not the first researchers making this distinction. Juola defines stylometry as something separate from authorship attribution, stating that stylometry is “determining any of the properties of the author(s) of a sample of text.” [4].

We are unable to find significant experimentation in the machine learning community targeting affiliation a level above the individual authorship attribution. Surprisingly, our extensive literature survey found only a handful of other peer reviewed research projects in the area of group stylometrics. In Juola’s recent survey, some of the ‘authorship properties’ he suggests that can be analyzed with stylometrics include “Was the author a native speaker of English? Of US or British English? A man or a woman?” [4]. Despite making these stylometric studies sound commonplace, Juola only specifically references a single paper addressing author gender, and within “formal written documents” at that [22], unlike the informally written website forum posts we are targeting.

Our hypothesis is that there are linguistic and grammatical nuances that are shared amongst people of a common background that can be detected within text communications, especially within the type of open, community environment that a forum tends to foster [1]. Our intuition is bolstered by the previously referenced gender classification study [22], a similar study which focused on classifying authors by age [23], and a study that focused on both [24]. The latter study achieved accuracy of ~80-90% using only two features: sentence length and slang word usage.

It is interesting to note that all previous studies focus on inherent traits that are not a conscious choice of the author, such as native language [29], unlike the group affiliation which we are targeting. Additionally, most of these methods have a heavy reliance on term frequency. We infer that these techniques work quite well [5] when the author’s topic is unconstrained, such as a novelist or blogger expounding on whatever suits their fancy. A secondary theory we will test is that in certain circumstances, specifically the ones we are targeting where the authors may be discussing extremely similar or identical subjects to each other. For example, they may be discussing a subject or context set *a priori*, such as responding to a previous post or posting within a specific forum topical section. In such cases, we speculate term frequency based approaches may not be sufficient for accurate classification, or may allow room for improvement. Thus our investigation includes both TF metrics and non-TF metrics.

II. Feature Extraction

We extracted 80 total linguistic features of 9 different categories. Some metrics were tracking authors word choice, which fall under the classic lexical & syntactic categories [8]: use of exclusive terms [1][18], negation terms [1][18], causation terms, function terms [3], pronoun usage [1][18], parts of speech [20], and method of noun referencing [1][19]. Other metrics track authors structural decisions and are considered syntactic/semantic [8]: Sentence length [24], and sentence closure. All of these features were also used in a recently by Khosmood to try to camouflage authorship through transformation [7]. The linguistic features are represented as a collection of ratios. Most of the ratios are expressed as a percentage of tokens within the document (for example, causation terms), but some are expressed as a relative to each other (for example, sentence length).

TABLE I. Linguistic Features

Feature Type	Linguistic Feature Definition
Exclusive Terms	But, except, without, exclude
Negation Terms	No, never, not, n’t
Causation Terms	Due, because, as, since, consequently, hence, so, therefore, accordingly, thus, if, unless, lest
Function Terms	303 specific terms from [9].
Noun Referencing	Definite/Indefinite Article/Demonstrative Pronoun immediately preceding noun
Sentence Length	% of Sentences of Length <10 words, 10-19 words, 20-29 words, 30-39 words, 40-49 words, and Length > 50 words.
Sentence Closure	Sentences / Questions / Exclamations
Parts of Speech	Ratio of terms classified in each of 50 different POS tag categories
Pronoun usage	Ratio of pronouns in each of 9 different categories: 2nd person, 3rd person, demonstrative, possessive, etc.

Our feature set includes a vector of counts of terms in a pre-defined dictionary depicting the *term-frequency* of that document. We used the same bag-of-words representation that is common in many authorship attribution studies as well as topic based text classification [10]. We chose the top-2000 most common terms as the dictionary for our bag-of-words representation, ignoring the stopwords in the default python NLTK stopwords list, which is a copy of the stopwords from the Cornell SMART system [17]. We used the default python POS tagger, a maximum entropy tagger trained against the University of Pennsylvania Treebank data set [30]. When using this feature representation style, it is common practice to normalize features, making them independent of document length. This step is especially important for us, because our application domain consists of informally written blog entries that vary from 1-2 sentences to 100’s of sentences. There are many ways to accomplish this document normalization, and we chose to use simple maximum-TF normalization scheme:

$$x_i = \frac{x_i}{\max(\vec{x})}. \quad (1)$$

We evaluated numerous other types of normalization popular in NLP literature (L2, L1, Augmented norm etc.) but did not observe notable performance difference in our tasks. This may be a property of our experimental set up and/or our data which allowed us to settle for this rather straightforward but effective normalization scheme.

III. Experimental Data Sets

We used two different data sets for our experiments. We selected one corpus to use as a baseline, and we created a second corpus of data representative of our problem domain.

The first is the BitterLemons.org corpus. This corpus was released in 2006 [2] and has been used in various linguistic studies since, making it a good baseline for our work. The corpus itself consists of 594 essays from an Israeli/Palestinian political discussion website (BitterLemons.org). Essays are published in sets of four on a particular topic. Each topic is addressed by an article from the same Israeli and Palestinian editors each week, as well as a guest Israeli and Palestinian commenter. In addition to its use within the ML/NLP community, there are two other characteristics which make this dataset particularly well suited to our experiment. First, the tightly balanced set of topics should minimize the amount of term/response bias as opposed to a traditional open website forum where authors post on whatever is of interest to them. Second, the corpus has a mixture of articles from various authors and styles, but has 'ground truth' very clearly labeled as to group affiliation (the guest commenters are vetted by the editors). Each article is clearly designated as to whether the author is representing the Israeli or the Palestinian side of the issue. We assume these labels to be noise-free ground truth in our experimentation.

The second corpus we used consisted of translated posts from website forums, the collection and translation of which was performed by intelligence analysts from another organization. The corpus consisted of two sets of articles. The first group consisted of 2636 Arabic language forum posts from 9 different website forums. These posts were originally written in Arabic, but we performed our experiments on an English language translation of the content. While we did not review every post in our data set individually, most of the forum posts in this set would likely be considered extremist from the current United States Government point of view.

For the other group in this data set, we collected 537 other forum posts. We collected these posts from 9 different websites that would most likely be considered 'moderate' from the current United States Government perspective. The forums we selected for this group included sites in Vietnam, Fiji, London, Afghanistan, and Bahrain. Again, we did not review every post individually but most of the content of these posts tended to focus on peaceful & diplomatic solutions. Most were written in Arabic, with a small number written in Vietnamese or other languages. All posts we targeted were also translated to English so that would be one less variable between data sets.

We use the classification of the top level domain of the URL (the website) of each forum post (extremist/moderate) as the label for group affiliation for our experimentation. This second data set has a ground truthing issue. Simply because a post was collected from a particular URL does not guarantee the author would have affinity for or identify with the group label that we are assigning the document. However fully accurate ground truth is unobtainable because it is impossible to read the mind of the original anonymous authors. In addition, unlike many internet based experiments where documents are collected or scraped randomly, in our case a trained intelligence analyst skimmed a larger portion of the website and designated our posts specifically for translation and inclusion of the document indicating its significance or that it reflected the mood of the site. So each post was reviewed as to its cohesiveness to the group, even if it was not by us specifically. Therefore we feel that this data set more than adequately reflects the scope of the problem, and is one of the more precise data sets available at this scale.

The other small issue is that we are processing the English translations of these documents, which loses some of the nuance pertaining to the Arabic language. However, these low-level features specifically pertaining to Arabic have already been used in experimentation by researchers [6]. Abbasi & Chen found that the "Key difference between English & Arabic feature sets" included word length, number of technical structures, number of elongations, and number of word roots (the last two being Arabic specific features). We did not have similar features in our testing, so the features we tested should hold across languages.

Also, it should be noted that our labels of the data are irrelevant, that is, which data set we would consider 'extremist'. The fact is that we selected distinct communities that could potentially have some overlap in vocabulary and structure, and are attempting to train our machine learning algorithms to discriminate between them.

IV. Machine Learning Experimentation

To test our hypothesis, we selected two of the most widely used and best performing supervised learning algorithms for document classification. Our goal is to generate a range of experiments to ensure our generalizations about the linguistic features are not limited to a particular classification methodology and/or a specific performance metric, as well as determine the feasibility of reliably performing author group affiliation.

A. *k-Nearest Neighbor (k-NN)*

k-Nearest Neighbor algorithm [11] is one of the simplest and most intuitive methods for classification. We utilized the standard method whereby a test input is classified as belonging to the majority category of its k nearest neighbors within a known labeled training set. Despite its simplicity, k-NN is known to yield competitive results when combined with a suitable metric for similarity (dissimilarity). In our experiments we selected the cosine distance as the metric to determine the similarity/dissimilarity between any given pair of examples. Cosine distance between two vectors x and y is given by:

$$d_{\cos}(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|} \qquad (2)$$

where $\| n \|$ is the l_2 norm of *n*. In the case of document matching, where the inputs are usually term-frequency vectors of documents to be compared, the cosine distance can be seen as a normalized distance measure independent of the effects of document length. This property of cosine distance makes it a popular metric in Information Retrieval (IR) applications. For completeness, we evaluated Euclidian distance in our k-NN algorithms and saw consistently lower performance than cosine distance. These results are omitted.

B. Support Vector Machines (SVM)

Support vector machines [12] are arguably one of the most successful and reliable classification algorithms in a supervised learning setting. They have been used extensively in many areas of machine learning including computer vision, speech recognition, network security analysis, and natural language processing (NLP). SVMs are linear classifiers by default. This can be a limitation with data which might not be linearly separable. To overcome this issue, SVMs are often used with kernel functions that map input data to higher (possibly infinite) dimensional feature space. This is commonly known in the machine learning community as the kernel trick [13], and this allows SVMs to learn highly non-linear boundaries in the original input feature space. One of the most widely used kernel functions is the Radial Basis Function kernel [12] (also referred to as the Gaussian kernel).

Another limitation often cited against the SVMs is its shallow learning architecture. Recently, Cho et al. introduced arc-cosine kernels that mimic the computations involved in neural networks [14]. These kernel functions provide a way to augment the SVM performance by exploiting the advantages of deep learning while keeping the learning problem convex and tractable unlike conventional deep learning architectures.

In this paper, we have experimented and compared the performances of linear SVMs, SVM with RBF kernels and SVMs with single and multi-layer arc-cosine kernels. For the linear and non-linear versions we used the libSVM library [15]. For our experiments with arc-cosine kernels we adapted the libSVM modification available online[1] to libSVM's Matlab interface. In addition, the libSVM library offers provisions to handle learning with heavily unbalanced classes while ensuring unbiased classifier training. Our second data set is heavily unbalanced because extremist posts are of more interest to analysts and are collected with more frequency.

C. Performance Metrics

To reduce the effect of an unbalanced dataset on our evaluations and to present a comparison that is independent of specific performance evaluators, we have compared our results under 3 different metrics: mean F-score, Matthew's Correlation Coefficient (MCC), and Balanced ACcuracy (BAC).

Mean F-score, as the name suggests, is the average of the F-scores corresponding to all the classes in the classification problem. The mean F-score of a binary classification with classes denoted as +1 and -1 can be represented as:

$$\mathrm{F}_{\mathrm{mean}} = 0.5 \times \left[\frac{2 \cdot \mathrm{P}_1 \cdot \mathrm{R}_1}{\mathrm{P}_1 + \mathrm{R}_1} + \frac{2 \cdot \mathrm{P}_{-1} \cdot \mathrm{R}_{-1}}{\mathrm{P}_{-1} + \mathrm{R}_{-1}} \right] \qquad (3)$$

where P and R are precision and recall respectively, the subscript indicates which class was considered as the 'relevant' class.

Matthew's Correlation Coefficient [26] is a metric that takes into account all 4 values in the confusion matrix while assessing performance and returns a value between -1 and +1. Under this metric, a random classification is given a score of 0 and a perfect classification is represented by +1. It is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)+(TN+FP)(TN+FN)}} \qquad (4)$$

where TP, TN, FP, FN are true positive, false negative, false positive and false negative rates respectively.

Balanced Accuracy is another metric that can be used if classification of both relevant and non-relevant classes needs to be treated with equal importance. It is defined in terms of *sensitivity* and *specificity* which measure a method's ability to identify positive and negative results respectively. The expression for BAC in terms of the elements of confusion matrix is:

$$BAC = 0.5 \times \left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right] \qquad (5)$$

where the first and second term denote sensitivity and specificity respectively.

D. Data Set Segmentation

We divided both of our datasets into randomly selected 70/30 splits for training and testing respectively. Whenever a development (validation) set was required (e.g. for SVM parameter selection), 20% of training set was randomly selected and set aside for validation purposes. The above process was repeated and averaged over 10 and 25 runs in the case of the Moderate-Extremist forum data and BitterLemons.org datasets respectively. The BitterLemons.org dataset was used in more trials because of its smaller size.

The random selections are intended to ensure a good mixture of our available documents in each combination, and this should reduce the chances of a few outlier documents unduly influencing results.

V. Experimental Results

We evaluated the performance of the linguistic features for authorship affiliation with the following classification algorithms: k-NN, linear SVM, SVM with RBF kernel, and SVM with the new arc-cosine kernels. For the k-NN algorithm, the neighborhood size parameter k was selected by conducting leave-one-out validation within the training set. The k-value that gave the best performance on each training set was used on

[1] available at http://cseweb.ucsd.edu/~yoc002/arccos.html

its corresponding test split. However, for SVMs, a separate validation set is necessary to avoid over-training. A simple grid search for parameters C and gamma (for RBF kernels) was performed as recommended by Hsu et al. [15], and the parameters that gave the best classification results on the set aside validation set were used on the corresponding test split. In the case of arc-cosine kernels, the degree coefficient and number of layers which can be varied were also treated as parameters to be tuned. However, on the validation datasets, we observed that a single layer of 0-th degree kernel often performed as well as or better than the higher degree kernels and multiple layers. This observation is in accordance with the findings of [16] which noted that single layer kernels have a tendency to outperform multiple layer kernels (possibly due to overfitting) when the features by themselves have enough discriminative ability [16]. Due to this reason, we have only presented results from single layer arc-cosine kernels of zero-th degree in this paper.

We have presented results obtained using regular term-frequency (TF), features identified by linguists (LIN) and a concatenated set of these features (TF+LIN). Although this is the simplest way to combine different types of features, it is important to note that even this naïve method is able to extract some performance boost.

The performance results obtained on our baseline dataset, entries from BitterLemons.org, is presented in Table III. As noted earlier, this dataset contains entries written in a controlled setting with 2 entries from each group on a particular topic. Despite the balanced and semi-formal nature of these entries, all 3 feature-sets show competitive discrimination between the groups. The TF and TF+LIN features with SVM provided scores ranging from 0.94-0.96 in both F-score and BAC and above 0.90 MCC score showing the clear distinction between the groups. The LIN feature scores, although lower than the TF scores, do prove to perform significantly higher than random indicating its suitability to identify distinctive styles used by different groups.

The classification results obtained on the moderate-extremist translated forum data is shown in Table II. This dataset has a wide range of stylistic differences between the two classes. This is captured by the linguistic style-based features. In terms of precision, recall, and Mean F-score, the LIN 80-dimensional feature set is able to perform comparable to the TF feature set which has a dimensionality of 2000. Also, the combined feature set attains a 3-6% improvement in k-NN performance whereas shows a steady improvement of slightly greater than 1% on all SVM methods. The greater improvement of k-NN is an important indication that by following a better methodology to combine features will enable SVMs also to capitalize on these added features.

Finally, the results from both these datasets reveal interesting trends. The fact that the concatenated feature set is able to consistently extract improvements under all performance metric indicates possible gains to be realized by better combining these different types of features. Also, the high performance of the classification algorithms on both the datasets with completely different group biases is promising. This indicates that by developing more linguistically oriented features and intelligently combining them with conventional TF-based features better and more reliable group detection can be performed.

VI. Conclusions

We were pleased with the high accuracy that we were able to achieve with our straightforward stylistic and TF features on our primary dataset (Table II) on our target forum data. In some cases this was higher than our benchmark BitterLemons dataset (Table III). While the inclusion of our 'linguistic' (LIN) features yielded a minimal boost, we feel that this is at least partially attributed to the extremely high performance of TF by itself on our particular test data. Yet the boost was consistently observed. In addition, when considering that the dimensionality of TF (2000) was an order of magnitude greater than LIN (80), coupled with the scale of our application area, seemingly marginal improvements in document classification could result in a savings of thousands of hours of intelligence analyst manpower. To that end, we feel that we have adequately demonstrated a performance gain following the inclusion of a small number of linguistic features compared to TF alone. Regardless, with our simplified feature combination scheme (concatenation), our results, both TF and augmented TF (TF+LIN) are extremely

TABLE II. Performance Comparison of regular TF, Linguistic Features and their combination on Moderate-Extremist Forum data

Metrics	**Mean F-score**			**MCC**			**Balance Accuracy**		
Features	*TF*	*LIN*	*TF+LIN*	*TF*	*LIN*	*TF+LIN*	*TF*	*LIN*	*TF+LIN*
k-NN	0.92	0.70	**0.95**	0.85	0.50	**0.91**	0.92	0.72	**0.95**
Linear SVM	0.96	0.77	**0.97**	0.92	0.60	**0.93**	0.96	0.78	**0.97**
SVM + RBF	0.96	0.80	**0.97**	0.92	0.64	**0.94**	0.96	0.81	**0.97**
SVM + Arccos	0.96	0.81	**0.97**	0.93	0.67	**0.94**	0.96	0.82	**0.97**

TABLE III. Performance Comparison of regular TF, Linguistic Features and their combination on BitterLemons data

Metrics	**Mean F-score**			**MCC**			**Balance Accuracy**		
Features	*TF*	*LIN*	*TF+LIN*	*TF*	*LIN*	*TF+LIN*	*TF*	*LIN*	*TF+LIN*
k-NN	0.89	0.76	**0.90**	0.79	0.53	**0.81**	0.89	0.76	**0.90**
Linear SVM	**0.95**	0.80	**0.95**	0.90	0.60	**0.91**	**0.95**	0.80	**0.95**
SVM + RBF	**0.95**	0.81	0.94	**0.90**	0.62	0.88	**0.95**	0.81	0.94
SVM + Arccos	**0.96**	0.82	0.95	**0.92**	0.64	0.90	**0.96**	0.82	0.95

encouraging. Considering the lack of existing research in the area of discerning authorship affiliation via any machine learning methodology, we have attempted to consider as many aspects as possible with respect to kernels and methods old and new. We feel there are numerous applications of this technology besides our own Department of Defense focus.

Future directions for research include better feature combination schemes including multiple kernel learning (MKL) [27], SVM-based discriminative accumulation scheme SVM-DAS [28], and other novel score weighting and feature selection methods. There are literally hundreds of other stylometrics that we could possibly include, but without appropriate filtering or combination methodologies, they just add noise and decrease performance.

Acknowledgment

The authors thank the Office of Naval Research and Martin Krueger for their support of this work. The authors also thank Dr. Marion Ceruti for her tireless contributions towards the fundamental research we applied within this experiment. This paper is the work of U.S. Government employees performed in the course of employment and no copyright subsists therein. This paper is approved for public release with an unlimited distribution.

References

[1] M.G. Ceruti, S.C. McGirr, and J.L. Kaina, “Interaction of Language, Culture and Cognition in Group Dynamics for Understanding the Adversary,” Proceedings of the National Symposium on Sensor and Data Fusion (NSSDF), 26-30 July, 2010 Nellis AFB, Las Vegas, NV. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA526247&Location=U2&doc=GetTRDoc.pdf

[2] W-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, “Which side are you on? Identifying perspectives at the document and sentence levels,” Proceedings of Tenth Conference on Natural Language Learning (CoNLL), pp. 109-116, 2006.

[3] R. Zheng, J. Li, H. Chen, Z. Huang, “A framework for authorship identification of online messages: Writing-style features and classification techniques” Journal of the American Society for Information Science and Technology Volume 57, Issue 3, pages 378–393, 1 February 2006

[4] P. Juola, “Authorship attribution”, Found. Trends Inf. Retr. 1, 3 pp. 233-334, December 2006.,

[5] G. Hirst and O. Feiguina, “Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts”, Literary & Linguist Computing 22(4): 405-417 first published online October 1, 2007 doi:10.1093/llc/fqm023

[6] A. Abbasi and H. Chen, “Applying Authorship Analysis to Extremist-Group Web Forum Messages”, presented at IEEE Intelligent Systems, pp.67-75, 2005.

[7] F. Khosmood, R. Levinson, “Automatic Synonym and Phrase Replacement Show Promise for Style Transformation” 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 958-961, 2010

[8] E. Stamatatos, “A survey of modern authorship attribution methods”, presented at JASIST, pp.538-556, 2009.

[9] S. Argamon, M. Saric, and S.S. Stein, “Style mining of electronic messages for multiple authorship discrimination: first results”, in Proceedings. KDD, pp.475-480, 2003

[10] F. Sebastiani, “Machine learning in automated text categorization”, presented at ACM Comput. Surv., pp.1-47., 2002

[11] T. Cover and P. Hart, “Nearest neighbor pattern classification”, In IEEE Transactions in Information Theory, IT-13, pages 21–27, 1967.

[12] C. Cortes and V. Vapnik, “Support-vector networks”, Machine learning, 20(3):273–297, 1995

[13] B. Schölkopf and A. J. Smola, “Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond” MIT Press, Cambridge, MA, 2001

[14] Y. Cho and L. Saul “Kernel methods for deep learning”, In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., & Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22, (pp. 342–350)., MIT Press, Cambridge, MA, 2009

[15] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, “A practical guide to support vector classification”, Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[16] C-C. Cheng, B. Kingsbury, “Arccosine Kernels: Acoustic modeling with infinite neural networks”, ICASSP 2011

[17] C. Buckley, G. Salton, and J. Allan, “Automatic Retrieval With Locality Information Using SMART”, in Proc. TREC, pp.59-72, 1992.

[18] J.F. Burrows, “Word patterns and story shapes: The statistical analysis of narrative style”.,Literary and Linguistic Computing, 1987

[19] E. Stamatatos, N. Fakotakis, and G.K. Kokkinakis, “Automatic Text Categorization In Terms Of Genre, Author”, presented at Computational Linguistics, pp.471-495,2000

[20] S. Argamon-Engelson, M. Koppel, & G. Avneri, “Style-based text categorization: What newspaper am I reading?” In Proceedings of AAAI Workshop on Learning for Text Categorization (pp. 1–4), 1998.

[21] A. Abbasi and H. Chen, “Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace”, presented at ACM Trans. Inf. Syst., 2008.

[22] M. Koppel, S. Argamon, A. Shimoni, “Automatically Categorizing Written Texts by Author Gender”, Literary and Linguistic Computing 17(3). 2002.

[23] M. Izumi, T. Miura, I. Shioya, “Estimating the date of blog authors by CRF”, IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 249 – 252. Victoria, BC, Aug. 2007.

[24] S. Goswami, S. Sarkar, and M. Rustagi, “Stylometric Analysis of Bloggers' Age and Gender”, In Proceedings of the AAAI International Confere (ICWSM). 2009.

[25] D. I. Holmes, “Authorship Attribution”, Computers and the Humanities 28(2), pp 87-106. Springer Netherlands, 1994.

[26] B. W. Matthews, , “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”, Biochim. Biophys. Acta 405,pp. 442–451, 1975.

[27] S. Sonnenburg, G. Raetsch, C. Schaefer, & B. Scholkopf, “Large scale multiple kernel learning”, Journal of Machine Learning Research, 7, 1531–1565, 2006.

[28] A. Pronobis, O. M. Mozos, and B. Caputo, “SVM-based discriminative accumulation scheme for place recognition”, In Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA’08), pp. 522-529, Pasadena, CA, USA, May 2008.

[29] M. Koppel, J. Schler, and K. Zigdon, “Determining an author's native language by mining a text for errors”, in Proc. Knowledge Discovery in Data Mining (KDD), pp.624-628, 2005.

[30] A Ratnaparkhi, “A Maximum Entropy Model for Part-Of-Speech Tagging”, in Proc. Of the Emperical Methods in Natural Language Processing (1996), pp. 133-142