Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Informatica

On Authorship Attribution

Relatore:
Chiar.mo Prof.
DANILO MONTESI

Presentata da:
Gabriele Calarota

Correlatore:
Dott.
FLAVIO BERTINI

Sessione III
Anno Accademico 2019-2020

"When I was in college,
I wanted to be involved in things that would change the world"
Elon Musk

Sommario

Abstract

Contents

chapter 4

State of the art: Data collection and Techniques for Authorship Attribution

'*State of the Art is the frenetic and relentless pursuit of doing what its best at that time!*'
Da Anunciação Marco

There are different types of authorship attribution studies in the literature such as predicting the date of authorship of historical texts or text genre detection [55], [25]. Vast majority of previous works focuses on authorship identification by taking into consideration the stylistic features of authors such as use of grammar, function words, frequent word allocations [2], [13], [18]. Some of the well-known problems in authorship attribution are disputed Federalist Papers classification and Shakespearean Authorship Dispute. The Federalist Papers are a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay to persuade the citizens of New York to ratify the U.S. Constitution. Authorship of twelve of these papers has been in dispute. To address this problem, using linear support vector machines as classifier and relative frequencies of words as features a study identified these papers to be written by James Madison [43].

Another dispute in authorship attribution among scholars across the world is whether William Shakespeare wrote the works attributed to him or not. It was argued that Shakespeare was not even educated and more than 80 authors were suggested to be the author of the writings that were under the name of Shakespeare. Christopher Marlowe is

considered the most likely candidate to write these works under the name of Shakespeare when he was in jail. In order to analyze the stylistic fingerprint of Shakespeare and Marlowe and non-Shakespearean authors, namely Chapman,Jonson, Middleton, a corpus has been put together [13].
The classification results for non-Shakespearean author candidates turned out to be highly accurate (Johnson %100, Chapman %92.9 and Middleton %88.9). The results supported the hypothesis that writing styles of Marlowe and Shakespeare were as distinguishable as other authors unless Marlowe did not show a linear change in style over time. Meaning, Marlowe has found not to be the authors of Shakespearean writings. Another interesting study on the unknown texts is also done based on word-level features, vocabulary richness and syntactic features by using Liblinear SVM for classification purposes [54]. Even though the classification accuracy results are not as high as other related works features like 'number of unique words' should be noted for use in any attribution problem.
Usefulness of function words in authorship attribution is introduced by Mosteller and Wallace in their work on Federalist papers [43]. Argamon and Levitan has compared the characteristic features of frequent words, pairs and collocations using the SMO algorithm, and implemented it for two class (American or British) author nationality classification problem. Their results conclude that function words are useful as stylistic text attribution and frequent words are the best features among others. The reason behind it is that a given same size frequent collocations has less different words comparing to frequent words so it carries less discriminatory features [2]. In summary, there has been substantial work done in authorship attribution and mainly people in forensic linguistic or computer scientists aim to build 'stylistic fingerprint of author' by using several features of a given text such as function words, stylometry. It is a classification problem and several classifiers are used such as Naïve Bayes, SVM. Among them, SVM is observed to fit best for these kinds of problems.

4.1 SVM

Support Vector Machines (SVMs) recently gained popularity in the learning community [58]. In its simplest linear form, an SVM is a hyperplane that separates a set of positive examples from a set of negative examples with maximum interclass distance, the margin. Figure 4.1 shows such a hyperplane with the associated margin. The formula for the output of a linear SVM is show in Equation 4.1, where w is the normal vector to the hyperplane, and x is the input vector. The margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples.

$$u = w * x + b \tag{4.1}$$

Maximizing the margin can be expressed as an optimization problem, as shown in Equation 4.2:

$$minimize \frac{1}{2}||w||^2 \; subject \; to \; y_i(w * x_i + b) \geq 1, \forall_i \tag{4.2}$$

where x_i is the $i-th$ training example and $y_i \in -1, 1$ is the correct output of the SVM for the $i-th$ training example. Note that the hyperplane is only determined by the training instances x_i on the margin, *the support vectors*. Support vector machines are based on the structural risk minimization principle from computational learning theory [58]. The idea is to find a model for which we can guarantee the lowest true error. This limits the probability that the model will make an error on an unseen and randomly selected test example. An SVM finds a model which minimizes (approximately) a bound on the true error by controlling the model complexity (VC-Dimension). This avoids over-fitting, which is the main problem for other semi-parametric models.

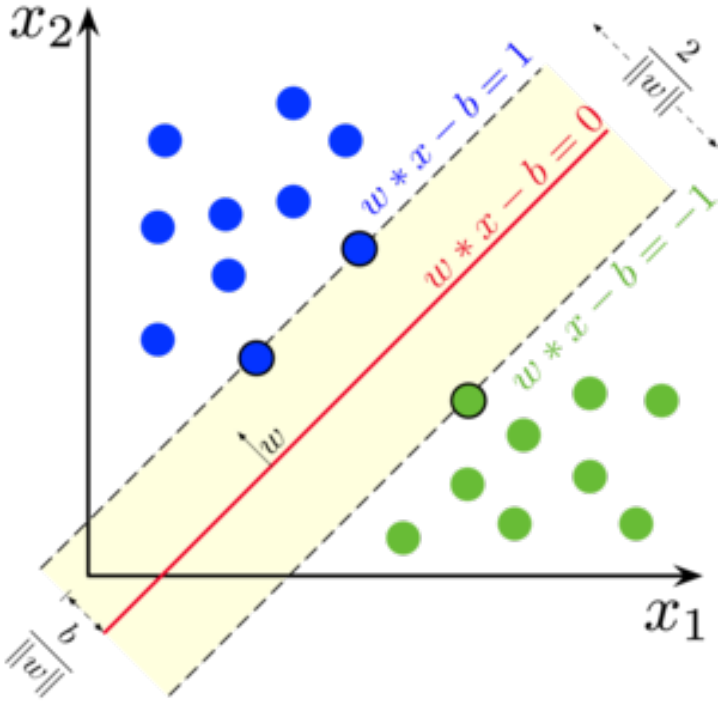

Figure 4.1: SVM Hyperplane with the associated margin formula

4.1.1 SVM for authorship attribution

Unlike currently used classification approaches, like neural networks or decision trees, SVM allows for the processing of hundreds of thousands of features. This offers the opportunity to use all words of a text as inputs instead of a few hundred carefully selected characteristic words only. In similar text classification problems aiming at thematic categorization, the SVM has been shown to be quite effective [22], [12]. A SVM is able to classify a text with respect to content. In the framework of author attribution, it is

not clear whether a specific topic addressed by the author or the structural or stylistic features of the authors language lead to a successful classification. Among the earliest methods to be applied were various types of neural networks, typically using small sets of FWs as features [20]. More recently, Hirst and Feiguina used neural networks on a wide variety of features. Other studies have used k-nearest neighbor [62], support vector machines [10], [28], [63], and Bayesian regression [35]. Comparative studies on machine learning methods for topic-based text categorization problems Dumais et al., Joachims have shown that in general, support vector machine (SVM) learning is at least as good for text categorization as any other learning method, and the same has been found for authorship attribution [63]. The distinctive advantage of the SVM for text categorization is its ability to process many thousand different inputs. This opens the opportunity to use all words in a text directly as features. For each word w_i the number of times of occurrence is recorded. Typically a corpus contains more than 100,000 different words, with each text covering only a small fraction. Joachims [22] used the SVM for the classification of text into different topic categories. As features he uses word stems. To establish statistically significant features he requires that each feature occurs at least three times in a text. The empirical evaluation was done on two test collections: the Reuter-21578 news agency data set covering different topics and the Ohsumed corpus of William Hersh describing diseases. Using about 10000 features in every case, the two SVM versions (polynomial and rbf) performed substantially better than the currently best performing conventional methods (naive Bayes, Rocchio, decision trees, k-nearest neighbor). Joachims et al. [23] used a transductive SVM for text categorization which is able to exploit the information in unlabeled training data. Dumais et al. [12] use linear SVMs for text categorization because they are both accurate and fast. They are 35 times faster to train than the next most accurate (a decision tree) of the tested classifiers. They applied SVMs to the Reuter-21578 collection, emails and web pages. Drucker et al. [11] classify emails as spam and non spam. They find that boosting trees and SVMs have similar performance in terms of accuracy and speed. SVMs train significantly faster.

4.2 RCV1 studies

Reuters is the world's largest international multimedia news agency, providing myriad news and mutual fund information available on `Reuters.com`, video, mobile, and interactive television platforms. Reuters Corpus Volume 1 (RCV1) is drawn from one of those online databases[1]. This dataset consists of all English language stories produced by Reuters journalists between August 20, 1996 and August 19, 1997. The dataset is made available on two CD-ROMs and has been formatted in XML by Reuters, Ltd. in 2000, for research

[1]Reuters corpora [Online]. Available, `http://trec.nist.gov/data/reuters/reuters.html`; 2000

purposes. Both the archiving process and later preparation of the XML dataset involved substantial verification and validation of the content, attempts to remove spurious or duplicated documents, normalization of dateline and byline formats, addition of copyright statements, and so on [8]. The stories cover a range of content typical of a large English language international newswire. They vary from a few hundred to several thousand words in length.

It consists of a collection of newswire stories written in English that cover four main topics: corporate/industrial (CCAT), economics (ECAT), government/social (GCAT) and markets (MCAT). Although it was not compiled for authorship attribution task, it has been adapted to this task in previous works. For example, in Stamatatos (2008); Plakias and Stamatatos (2008) the 10 most prolific authors were chosen from the CCAT category, and then, 50 examples per author for training and 50 examples for testing were selected randomly with no overlapping between training and testing sets. In further sections, we will reference to this corpus as RCV1-10.

In Houvardas and Stamatatos [21], the authors proposed another adaptation of the RCV1 corpus for the authorship attribution task. They choose the 50 most prolific authors from the Reuters Corpus, keeping 50 examples per author for training and 50 examples per author for testing with no overlapping between them. We will refer to this corpus as RCV1-50.

The RCV1-10 and RCV1-50 datasets are both balanced over different authors and have their genre fixed to news. The main category of the news in both cases is fixed to corporate/industrial, but there are many subtopics covered in the news and the length of the texts is short (from 2 to 8 KBytes). These corpora resemble a more realistic scenario, when the amount of texts is limited and the number of candidate authors is large.

4.2.1 Studies on RCV1 on authorship attribution

Although, not particularly designed for evaluating author identification approaches, the RCV1 corpus contains 'by-lines' in many documents indicating authorship. In particular, there are 109,433 texts with indicated authorship and 2,361 different authors in total. RCV1 texts are short (approximately 2KBytes – 8KBytes), so they resemble a realworld author identification task where only short text samples per author may be available. Moreover, all the texts belong to the same text genre (newswire stories), so the genre factor is reduced in distinguishing among the texts. On the other hand, there are many duplicates (exactly the same or plagiarized texts). The RCV1 corpus has already been used in author identification experiments. In Khmelev and Teahan [27] the top 50 authors (with respect to total size of articles) were selected. Moreover, in the framework of the AuthorID project, the top 114 authors of RCV1 with at least 200 available text samples were selected [35]. In contrast to these approaches, in this study, the criterion for selecting

the authors was the topic of the available text samples. Hence, the top 50 authors of texts labeled with at least one subtopic of the class CCAT (corporate/industrial) were selected. That way, it is attempted to minimize the topic factor in distinguishing among the texts. Therefore, since steps to reduce the impact of genre have been taken, it is to be hoped that authorship differences will be a more significant factor in differentiating the texts. Consequently, it is more difficult to distinguish among authors when all the text samples deal with similar topics rather than when some authors deal mainly with economics, others with foreign affairs etc. The training corpus consists of 2,500 texts (50 per author) and the test corpus includes other 2,500 texts (50 per author) non-overlapping with the training texts [21].

4.3 GDELT studies

The GDELT Project is one of the largest publicly available digitized book database which has more than 3.5 million books published from 1800-2015. The GDELT Project is an open platform for research and analysis of global society and thus all datasets released by the GDELT Project are available for unlimited and unrestricted use for any academic, commercial, or governmental use of any kind without any fee[2]. The whole digitized dataset is publicly available and interested researchers can freely perform SQL queries using the Google big query platform. For example; the book names, publication year, quotations, themes, the original text of the book of "Mark Twain" which were written between 1890 to 1900 can be found as follows using the Big query platform of Google in Code Listing 4.1.

```
SELECT Themes, V2Themes, Quotations, AllNames, TranslationInfo,
BookMeta_Identifier, BookMeta_Title, BookMeta_Creator,
BookMeta_Subjects, BookMeta_Year,
FROM (TABLE_QUERY([gdelt-bq:internetarchivebooks],
'REGEXP_EXTRACT(table_id, r"(\d{4})") BETWEEN "1890" AND "1900"'))
WHERE
BookMeta_Creator CONTAINS "Mark Twain"
LIMIT 50
```

Code Listing 4.1: Google Big Query on GDELT

To decrease the bias and create a reliable dataset the following criteria have been chosen to filter out authors: English language writing authors, authors that have enough books available (at least 5), 19th century authors. With these criteria 50 authors have been selected and their books were queried through Big Query Gdelt database. The next task

[2] https://www.gdeltproject.org/about.html

has been cleaning the dataset due to OCR reading problems in the original raw form. To achieve that, firstly all books have been scanned through to get the overall number of unique words and each words frequencies. While scanning the texts, the first 500 words and the last 500 words have been removed to take out specific features such as the name of the author, the name of the book and other word specific features that could make the classification task easier. After this step, we have chosen top 10, 000 words that occurred in the whole 50 authors text data corpus. The words that are not in top 10, 000 words were removed while keeping the rest of the sentence structure intact. Afterwards, the words are represented with numbers from 1 to 10, 000 reverse ordered according to their frequencies. The entire book is split into text fragments with 1000 words each. We separately maintained author and book identification number for each one of them in different arrays. Text segments with less than 1000 words were filled with zeros to keep them in the dataset as well. 1000 words make approximately 2 pages of writing, which is long enough to extract a variety of features from the document. The reason why we have represented top 10, 000 words with numbers is to keep the anonymity of texts and allow researchers to run feature extraction techniques faster. Dealing with large amounts of text data can be more challenging than numerical data for some feature extraction techniques.

4.3.1 Authorship attribution GDELT

4.4 The guardian studies

First introduced by [53] The corpus used in this study is composed of texts published in The Guardian daily newspaper. The texts were downloaded using the publicly available API20 and preprocessed to keep the unformatted main text.21 An example is depicted in Table 1. The majority of the corpus comprises opinion articles (comments). The newspaper describes the opinion articles using a set of tags indicating its subject. There are eight top-level tags (World, U.S., U.K., Belief, Culture, Life&Style, Politics, Society), each one of them having multiple subtags. It is possible (and very common) for an article to be described by multiple tags belonging to different main categories (e.g., a specific article may simultaneously belong to U.K., Politics, and Society). In order to have a clearer picture of the thematic area of the collected texts, we only used articles that belong to a single main category. Therefore, each article can be described by multiple tags, all of them belonging to a single main category. Moreover, articles coauthored by multiple authors were discarded. In addition to opinion articles on several thematic areas, the presented corpus comprises a second text genre—book reviews. The book reviews are also described by a set of tags similar to the opinion articles. However, no thematic tag restriction was taken into account when collecting book reviews, since our

main concern was to find texts of a specific genre that cover multiple thematic areas. Note that since all texts come from the same newspaper, they are expected to have been edited according to the same rules, so any significant difference among the texts is not likely to be attributed to the editing process. Table 1 shows details about The Guardian Corpus ("TGC"). It comprises texts from thirteen authors selected on the basis of having published texts in multiple thematic areas (Politics, Society, World, U.K.) and different genres (opinion articles and book reviews). At most 100 texts per author and category have been collected—all of them published within a decade (from 1999 to 2009). Note that the opinion article thematic areas can be divided into two pairs of low similarity, namely PoliticsSociety and World-U.K. In other words, the Politics texts are more likely to have some thematic similarities with World or U.K. texts than with the Society texts. TGC provides texts on two different genres from the same set of authors. Moreover, one genre is divided into four thematic areas. Therefore, it can be used to examine authorship attribution models under cross-genre and cross-topic conditions. [15] Various techniques were developed for solving the AA task. In [24], the authors present a reproducibility study on AA research. They evaluated state-of-the-art methods on three corpora and showed that only four out of fifteen (4/15) approaches are stable across corpora. The majority of the studies on AA perform extensive feature engineering, focusing on the extraction of stylometric features that represent the personal style of authors [7,25,28]. The approaches that employ character-based features (character n-grams) seem to be the most effective ones for the AA problem under both single and cross-topic conditions [5,25]. Previous work on AA focused mainly on the single-topic condition, i.e., the training and testing datasets have similar thematic properties. However, there are studies that tackle the AA problem under cross-topic conditions. Stamatatos [31] demonstrated that high frequency character n-grams allow to discriminate effectively between authors not only for single-topic AA, but also for cross-topic AA. The unmasking method yields reliable results under cross-genre [10] and cross-topic conditions [12]. Sapkota et al. [26] improve the prediction results in cross-topic AA using an enriched training corpus in order to predict authors on a corpus with different topics. The role of preprocessing steps was evaluated in [18]. The approach proposed in that paper is considered to be more topic-neutral by their authors, because they replace the named entities and some topic-related words while preprocessing the corpus. Their approach showed the importance of preprocessing, because it gave the improvement of 4%. In [30], it is mentioned that the use of semantic features for the authorship attribution task usually improves the obtained results, however, very few attempts have been done to exploit high-level features for stylometric purposes. In this paper, we consider the usage of the distributed document representation for the cross-topic AA task, because of its capability to encode the semantic information of texts in a low dimension vector. Different document embeddings methods

were introduced in recent studies, each tackling specific natural language processing tasks: weighted concatenation of word vectors [16], deep averaging network [8], paragraph vector n-gram model [15], recurrent neural networks [29], and convolutional neural networks [9]. Recently, the Paragraph Vector (Doc2vec) model was proposed by Le & Mikolov [14] for learning distributed representation for both sentences and documents. The Doc2vec model basically treats each document as a special word and learn both document vectors and word vectors simultaneously by predicting the target word. Vectors obtained by the Doc2vec model outperforms both bag-of-words and word ngrams models producing the new state-of-the-art results for several text classification and sentiment analysis tasks. The experiments were conducted on a cross-topic AA corpus introduced in [31]. The corpus contains a set of articles gathered from 1999 to 2009 from the English newspaper The Guardian. 2 The articles corresponding to 13 authors were collected and grouped into five topic categories: Politics, Society, World, UK, and Book reviews. In order to avoid category overlapping, those articles whose content includes more than one category were discarded. In this way, each category is mutually exclusive. An important remark about the proposed categories is that all the categories are composed by articles, but the Book reviews are considered different text genre. The Guardian corpus is a specific collection of samples, which provides both a cross-topic scenario (five different topics) and a cross-genre scenario (articles and reviews) for AA task. The number of samples in The Guardian corpus is not balanced. The gathered samples correspond to a realistic scenario that considers the production of each author over a period of 10 years. In order to test and compare our approach, we reproduce the testing scenario described in the previous research [31] using the Guardian corpus. The experimental scenario is as follows: (1) select at most ten samples per author in each topic category (in Fig. 2 the distribution of the samples per author and per category after considering the restriction of ten samples per author is shown), (2) use the samples in the Politics category as training set and train the classifier, and (3) finally, test the classifier using another topic category different from Politics (four possible pairings).

4.4.1 Cross-topic authorship attribution

4.5 Studies on Stanford Amazon Food Reviews

4.6 Dataset selection

Bibliography

[1] Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[2] Shlomo Argamon and Shlomo Levitan. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, pages 4–7, 2005.

[3] Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. An experiment in authorship attribution. In *6th JADT*, volume 1, pages 69–75. Citeseer, 2002.

[4] Sarkhan Badirli, Mary Borgo Ton, Abdulmecit Gungor, and Murat Dundar. Open set authorship attribution toward demystifying victorian periodicals. *arXiv preprint arXiv:1912.08259*, 2019.

[5] John Burrows. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287, 2002.

[6] JK Chambers, P Trudgill, and Natalie Schilling-Estes. The handbook of language variation and change (2nd). *Victoria: Blackwell Publishing*, 2004.

[7] Carole E Chaski. Who's at the keyboard? authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1):1–13, 2005.

[8] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[9] Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, 1:343–359, 2007.

[10] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied intelligence*, 19(1):109–123, 2003.

[11] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural networks*, 10(5):1048–1054, 1999.

[12] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, 1998.

[13] Neal Fox, Omran Ehmoda, and Eugene Charniak. Statistical stylometrics and the marlowe-shakespeare authorship debate. *Proceedings of the Georgetown University Roundtable on Language and Linguistics (GURT), Washington, DC, USA*, 2012.

[14] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th international conference on Software engineering*, pages 893–896, 2006.

[15] Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and David Pinto. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):741–756, 2018.

[16] Jack Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3):251–270, 2007.

[17] H van Halteren. Linguistic profiling for authorship recognition and verification. 2004.

[18] Graeme Hirst and Ol'ga Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007.

[19] David I Holmes. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117, 1998.

[20] David I Holmes and Fiona J Tweedie. Forensic stylometry: A review of the cusum controversy. *Revue Informatique et Statistique dans les Sciences Humaines*, 31(1): 19–47, 1995.

[21] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 77–86. Springer, 2006.

[22] Thorsten Joachims. Making large-scale svm learning practical. Technical report, Technical report, 1998.

[23] Thorsten Joachims et al. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209, 1999.

[24] Patrick Juola. *Authorship attribution*, volume 3. Now Publishers Inc, 2008.

[25] Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic detection of text genre. *arXiv preprint cmp-lg/9707002*, 1997.

[26] Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the cross-domain authorship attribution task at pan 2019. In *CLEF (Working Notes)*, 2019.

[27] Dmitry V Khmelev and William J Teahan. A repetition based measure for verification of text collections and for text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 104–110, 2003.

[28] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628, 2005.

[29] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660, 2006.

[30] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26, 2009.

[31] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.

[32] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Yaron Winter. The "fundamental problem" of authorship attribution. *English Studies*, 93(3):284–291, 2012.

[33] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[34] David D Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, volume 33, pages 81–93, 1994.

[35] David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*, 2005.

[36] Yuval Marton, Ning Wu, and Lisa Hellerstein. On compression-based text classification. In *European Conference on Information Retrieval*, pages 300–314. Springer, 2005.

[37] S Michaelson and A Morton. The qsum plot. Technical report, Internal Report CSR-3, 1990.

[38] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[40] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[41] Leonid A Mironovsky, Alexander V Nikitin, Nina N Reshetnikova, and Nikolay V Soloviev. Graphological analysis and identification of handwritten texts. In *Computer Vision in Control Systems-4*, pages 11–40. Springer, 2018.

[42] Tom M Mitchell. Artificial neural networks. *Machine learning*, 45:81–127, 1997.

[43] Frederick Mosteller and David L Wallace. *Inference and disputed authorship: The Federalist*. Stanford Univ Center for the Study, 2007.

[44] Rebekah Overdorf and Rachel Greenstadt. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies*, 2016(3):155–171, 2016.

[45] Spyridon Plakias and Efstathios Stamatatos. Tensor space models for authorship identification. In *Hellenic Conference on Artificial Intelligence*, pages 239–249. Springer, 2008.

[46] Joseph Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365, 1997.

[47] Conrad Sanderson and Simon Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 482–491, 2006.

[48] Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 93–102, 2015.

[49] Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891, 2013.

[50] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[51] Efstathios Stamatatos. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2):790–799, 2008.

[52] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[53] Efstathios Stamatatos. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2):421–439, 2013.

[54] Sean Stanko, Devin Lu, and Irving Hsu. Whose book is it anyway? using machine learning to identify the author of unknown texts. *Machine Learning Final Projects*, 2013.

[55] Andrew Tausz. Predicting the date of authorship of historical texts. *CS224N project*, 2011.

[56] Antônio Theóphilo, Luís AM Pereira, and Anderson Rocha. A needle in a haystack? harnessing onomatopoeia and user-specific stylometrics for authorship attribution of micro-messages. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2692–2696. IEEE, 2019.

[57] Chris van der Lee and Antal van den Bosch. Exploring lexical and syntactic features for language variety identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 190–199, 2017.

[58] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[59] Carrington B Williams. Mendenhall's studies of word-length distribution in the works of shakespeare and bacon. *Biometrika*, 62(1):207–212, 1975.

[60] Lili Yang, Chunping Li, Qiang Ding, and Li Li. Combining lexical and semantic features for short text classification. *Procedia Computer Science*, 22:78–86, 2013.

[61] G Udny Yule. A test of tippett's random sampling numbers. *Journal of the Royal Statistical Society*, 101(1):167–172, 1938.

[62] Ying Zhao and Justin Zobel. Effective and scalable authorship attribution using function words. In *Asia Information Retrieval Symposium*, pages 174–189. Springer, 2005.

[63] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan 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*, 57(3): 378–393, 2006.

[64] George Kingsley Zipf. Selected studies of the principle of relative frequency in language. 1932.

[65] Sven Meyer Zu Eissen, Benno Stein, and Marion Kulig. Plagiarism detection without reference collections. In *Advances in data analysis*, pages 359–366. Springer, 2007.