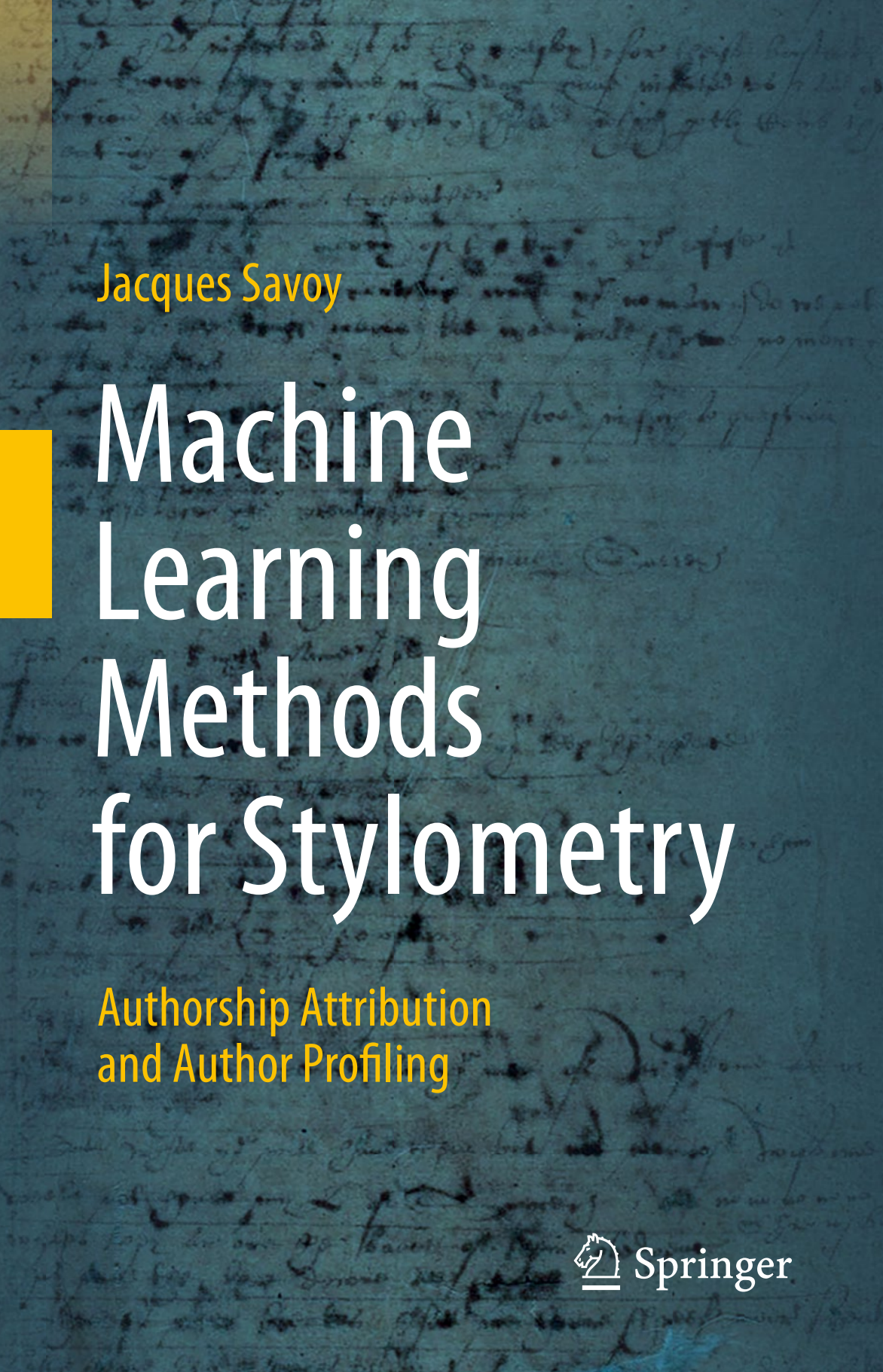
Jacques Savoy
Machine Learning Methods for Stylometry
Authorship Attribution and Author Profiling
Springer

Machine Learning Methods for Stylometry

Jacques Savoy

Machine Learning Methods for Stylometry

Authorship Attribution and Author Profiling

Springer

Jacques Savoy
Department of Computer Science
University of Neuchatel
Neuchâtel, Switzerland

ISBN 978-3-030-53359-5 ISBN 978-3-030-53360-1 (eBook)
https://doi.org/10.1007/978-3-030-53360-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Jacinthe, Adelaïde, and Benjamin

Preface

With the recent progress made in network and computing technology, the ubiquity of data, and textual repositories freely available, the scientific practice evolves towards a more data-based methodology. Thus, numerous domains consider machine learning models as pertinent tools to verify hypotheses or to improve their knowledge by discovering significant patterns hidden in datasets. And stylometry, or more generally digital humanities, follows this new research trend.

Focusing on the written style, this book presents methods and approaches able to identify the true author of a doubtful document or text excerpt. Assuming that each author has his[1] specific style, statistical or computer-based models can be applied to verify whether or not Shakespeare was the real author of a given play or poem. Besides literature works and authorship attribution, stylometric approaches can be useful to determine some demographics about the author. For example, one can wonder whether a novel (e.g., *My Brilliant Friend* (2012) by Elena Ferrante) is really written by a female writer. As other factors having a significant impact on the written style, one can study the effect of the author's age or his origin and native language. Instead of targeting the author, stylometric methods can be applied to draw the overall picture of style variations over a given time period or to underline the stylistic differences among a set of writers. With the ubiquity of social networks, stylometry can also be employed to infer some psychological traits of the author of a set of tweets as well as to identify early signs of depression. As a last example, stylometric measurements can be utilized to identify documents generated by machine or tweets sent by bots. This last aspect is related to the need to automatically detect fake news and its means and modes of dissemination.

Thus, the main intent of this book is to provide a broad introduction to all these text categorization problems grounded on stylistic features. This field of interest is clearly a multi-disciplinary one requiring some understanding in linguistics (or simply having some interest in this domain) and a basic knowledge in both statistics

[1] To simplify the presentation, the masculine form has been selected to indicate equivalently a man or a woman.

and computer science. We do not expect that the reader has an advanced skill in all these three domains. Thus, if the reader wishes to revise his knowledge in statistics, a gentle introduction, in plain English, is provided by Spiegelhalter [370]. As a solving-based approach is adopted in the different presentations of this book, explanations are supported by examples written in R. With S, its predecessor, this open-source software has radically changed the way statistics and data processing are applied; we moved from the pencil, calculator, and the use of various tables to statistical computing leading to modern data science. If the reader feels the need to acquire a better knowledge of R, an uncomplicated introduction is available in [41]. Other books expose the R software in a linguistics context (see [90, 182], or [139]).

Book Structure

This book is subdivided into three parts. The normal reading sequence follows the chapter order. However, depending on the interest of the reader, some chapters could be skipped in a first reading. More precisely, the first part presents a general introduction and some well-known models for solving the authorship attribution question. This section is dedicated to readers having a background in the humanities. The second part (Chaps. 4–7) is more devoted to computer science with a focus on machine learning models. The third part corresponding to the last three chapters presents real stylometric applications and can be read by everybody. As sequential reading is not mandatory, some redundancy will appear from time to time.

In more detail, the first part covering the first three chapters proposes a general introduction to the stylometry domain with its possible applications and limits. After describing the main factors explaining written styles, our running example used to illustrate the presented concepts is exposed. Various overall stylistic measurements are defined and commented upon. Finally, Chap. 3 presents the four most frequently used stylometric modes to solve authorship attribution problems in the humanities.

Chapters 4–7 form the second part. As a fair evaluation methodology is crucial, this section starts with a chapter on this question. As this chapter contains more statistical arguments, it could be omitted in a first reading. As the main aim of this second part is to explain machine learning models for solving stylometric problems, a chapter exposes several general strategies to identify, extract, select, and represent stylistic markers. As fundamental models, this section presents the k-nearest neighbors (k-NN), the naïve Bayes, the support vector machine (SVM), and the logistic regression and applies them to our running example. The last chapter presents more recent approaches proposed to solve the authorship problem (e.g., the Zeta test, compression, latent Dirichlet allocation (LDA)). In addition, more specific methods have been developed for providing answers to more specific questions such as the verification issue (Is Shakespeare the author of the play *The Tempest*?) or to detect possible joint collaboration to write a novel. As the deep learning approach represents an active field of research, a presentation of neural network models and

word embeddings applied to stylometry is provided as well as a general introduction to the deep learning approach to solve stylometric questions.

The last part embraces the last three chapters, each of them focusing on a particular question. The main intent of this last section is to illustrate with real cases the application of the different approaches. When needed, complementary information can be obtained by following references to previous passages in this book. As application, Chap. 8 presents an authorship attribution problem, to know who is the secret hand behind the *nom de plume* Elena Ferrante, an Italian writer worldwide known for her *My Brilliant Friend*'s saga. The second real case concerns social media and more particularly the social medium platform Twitter. The subject is to verify whether a computer can identify if a set of tweets have been generated by a bot or a human being, and in this second alternative, if it was written by a man or a woman. The last application exposes various strategies to explore stylistic variations over time using US political speeches covering a period of around 230 years.

Hands-On Exercises and Examples

To complement the presentation and discussion about stylometric models and techniques, examples and datasets are freely available. These illustrative examples are coded using the R software. This open-source language and the interpreter can be downloaded from the Internet at the following address:

`http://cran.r-project.org/`

It is important to know that the R software is used worldwide in statistics, both in academia and in industrial projects often related to big data applications. Knowing R is certainly a salient asset in your curriculum.

In addition, the datasets and the R code of our examples proposed in this book are freely available in the following GitHub webpage:

`https://github.com/JacquesSavoy/style`

For readers wishing to apply the presented methods on our examples or with other novels, we encourage the readers to download the `stylo` package written for R from the following URL:

`https://github.com/computationalstylistics/stylo`

This package provides useful stylometric functions and methods as well as additional novels to test them on. Our dedicated webpage also contains additional examples based on the `stylo` package to present advanced text representations or authorship models.

As a convention in this book, the `courier` font indicates statements, variables, or file names used in our examples written in R. *Italics* formatting is used when introducing important concepts but also to signal the novel or play titles and the occasional foreign terms. Finally, an index is available at the end of this book for a quick reference to the most important concepts.

Neuchâtel, Switzerland Jacques Savoy
May 2020

Acknowledgements

Without discussions and research done by colleges, this book would not have been possible. First, I want to mention Prof. Dominique Labbé (University of Grenoble) for introducing me to the stylometry questions, methods, and evaluation measures. I also thank Prof. Arjuna Tuzzi (University of Padova) for organizing several summer schools on quantitative analysis of textual data IQLA-GIAT in Padova. She was also an important contributor inside the IQLA (International Quantitative Linguistics Association) and in developing the Elena Ferrante corpus (see Chap. 8). Of course, I should also mention other main contributors of the IQLA, namely, and in alphabetical order, Prof. Maciej Eder (Polish Academy of Sciences, Kraków), Prof. Patrick Juola (Duquesne University, Pittsburgh), Prof. George Mikros (National and Kapodistrian University of Athens), and Prof. Jan Rybicki (Jagiellonian University, Kraków).

My recognition also goes to the various people in charge for organizing all the CLEF PAN evaluation campaigns, to mention a few of them (in alphabetical order), Prof. Shlomo Argamon (Illinois Institute of Technology, Chicago), Prof. Moshe Koppel (Bar-Ilan University), Prof. Martin Potthast (Leipzig University), Dr. Francisco Rangel (Symanto Research), Prof. Paolo Rosso (Universidad Politecnica de Valencia), and Prof. Efstathios Stamatatos (University of Aegean).

Contents

Part I Fundamental Concepts and Models

1 Introduction to Stylistic Models and Applications 3
- 1.1 Overview and Definitions 4
- 1.2 Style and Its Explaining Factors 5
- 1.3 Authorship Attribution 9
- 1.4 Author Profiling 10
- 1.5 Forensic Issues 13
- 1.6 Author Clustering 15
- 1.7 Other Related Problems 16

2 Basic Lexical Concepts and Measurements 19
- 2.1 Stylometric Model 20
- 2.2 Our Running Example: The *Federalist Papers* 21
- 2.3 The Zipf's Law 23
- 2.4 Vocabulary Richness Measures 26
- 2.5 Overall Stylistic Measures 30
- 2.6 And the Letters? 32

3 Distance-Based Approaches 33
- 3.1 Burrows' Delta 34
- 3.2 Kullback–Leibler Divergence Method 39
- 3.3 Labbé's Intertextual Distance 42
- 3.4 Other Distance Functions 44
- 3.5 Principal Component Analysis (PCA) 46

Part II Advanced Models and Evaluation

4 Evaluation Methodology and Test Corpora 55
- 4.1 Preliminary Remarks 55
- 4.2 Text Quality and Preprocessing 57
- 4.3 Performance Measures 59
- 4.4 Precision, Recall, and F1 Measurements 63

4.5 Confidence Interval 65
4.6 Statistical Assessment 67
4.7 Training and Test Sample 71
4.8 Classical Problems 73
4.9 CLEF PAN Test Collections 76
4.10 Evaluation Examples 78

5 Features Identification and Selection 83
5.1 Word-Based Stylistic Features 84
5.2 Other Stylistic Feature Extraction Strategies 87
5.3 Frequency-Based Feature Selection 93
5.4 Filter-Based Feature Selection 95
5.5 Wrapper Feature Selection 103
5.6 Characteristic Vocabulary 104

6 Machine Learning Models 109
6.1 *k*-Nearest Neighbors (*k*-NN) 110
6.2 Naïve Bayes 117
6.3 Support Vector Machines (SVMs) 123
6.4 Logistic Regression 131
6.5 Examples with R 136
6.5.1 *K*-Nearest Neighbors (*k*-NN) 136
6.5.2 Naïve Bayes 140
6.5.3 Support Vector Machines (SVMs) 145
6.5.4 Logistic Regression 148

7 Advanced Models for Stylometric Applications 153
7.1 Zeta Method 153
7.2 Compression Methods 157
7.3 Latent Dirichlet Allocation (LDA) 160
7.4 Verification Problem 162
7.5 Collaborative Authorship 168
7.6 Neural Network and Authorship Attribution 172
7.7 Distributed Language Representation 176
7.8 Deep Learning and Long Short-Term Memory (LSTM) 180
7.9 Adversarial Stylometry and Obfuscation 184

Part III Cases Studies

8 Elena Ferrante: A Case Study in Authorship Attribution 191
8.1 Corpus and Objectives 192
8.2 Stylistic Mapping of the Contemporary Italian Literature 195
8.3 Delta Model 198
8.4 Labbé's Intertextual Distance 202
8.5 Zeta Test 205
8.6 Qualitative Analysis 208
8.7 Conclusion 209

9 Author Profiling of Tweets 211
9.1 Corpus and Research Questions 212
9.2 Bots versus Humans 216
9.3 Man vs. Woman 219
9.4 Conclusion 227

10 Applications to Political Speeches 229
10.1 Corpus Selection and Description 230
10.2 Overall Measurements 232
10.3 Stylistic Similarities Between Presidencies 235
10.4 Characteristics Words and Sentences 240
10.5 Rhetoric and Style Analysis by Wordlists 243
10.6 Conclusion 249

11 Conclusion 251

Appendix A 255
A.1 Additional Resources and References 255
A.2 The Most Frequent Word-Types in the *Federalist Papers* 256
A.3 Proposed Features for the *Federalist Papers* 259
A.4 Feature Selection 260
A.5 Most Frequent Terms in Italian 262
A.6 US Presidents 263

References 265

Index 283

Acronyms

Many acronyms and abbreviations are used in this book. For the most frequent ones, the following list provides the corresponding full name and for some of them a short definition.

AA	Authorship attribution
ASCII	American Standard Code for Information Interchange
BW	Big word, word composed of six or more characters
CHI	Chi-square distribution
DNA	Deoxyribonucleic acid
FN	False negative
FP	False positive
FW	Functional words, corresponding to determiners, pronouns, conjunctions, prepositions, auxiliary and some modal verbal forms
GR	Gain ratio
HTML	HyperText Markup Language
IG	Information gain
LD	Lexical density, percentage of content-bearing words in a text
LDA	Latent Dirichlet allocation
LGBT	Lesbian, gay, bisexual, and transgender
LNRE	Large number of rare events
MeSH	Medical Subject Headings
MFL	Most frequent lemma
MFT	Most frequent word-type
MFW	Most frequent word, implicitly word-type
MSL	Mean sentence length
NN	Neural network
OR	Odds ratio
P	Precision
PMI	Pointwise mutual information
POS	Part-of-speech, or grammatical category or word class
QLF	Quadratic loss function

R	Recall
RNN	Recurrent neural network
RR	Reciprocal rank
SMS	Short message service
TC	Text categorization
TN	True negative
TP	True positive
TTR	Type–token ratio
URL	Uniform resource locator
WER	Word error rate
WWW	World Wide Web
XML	eXtensible Markup Language

List of Symbols

The following list indicates the main variables used in this book together with their definition. For example, the variable n indicates the number of tokens occurring in a text, a sample of word-texts, or in a corpus (depending on the context). To represent the absolute occurrence frequency of the ith term, the variable tf_i (where tf means *term frequency*) is used. Depending on the context, this absolute frequency is computed only according to a single document or according to the entire corpus. The notation $tf_{i,j}$ is employed to indicate the absolute frequency of the ith term in the jth text. The variable rtf_i denotes the *relative term frequency* of the ith term (with respect to a given document, author profile, or corpus).

ω	An arbitrary word-type
c	A constant
$Voc(T)$	The set of word-types appearing in the text T (vocabulary)
$\|Voc(T)\|$	The vocabulary size of text T
$Voc_k(T)$	The set of word-types appearing exactly k times in the text T
$\|Voc_k(T)\|$	The number of word-types appearing exactly k times in the text T
n	The number of tokens (in the text)
m	The number of stylistic features (terms) in a model
r	The number of possible categories (or classes, authors)
ln()	The natural logarithm (and log() is logarithm with a basis = 10)
$func(T)$	A function (to be specified) on the text T (e.g., length(T) = n)
$tf_{i,j}$	The absolute term frequency of the ith term in the jth text
tf_i	The absolute term frequency of the ith term
$rtf_{i,j}$	The relative term frequency of the ith term in the jth text
rtf_i	The relative term frequency of the ith term
df_i	The absolute document frequency of the ith term
r_i	The ith rank
t_i	The ith term
$t_{i,j}$	The ith term in the jth text
D_j	The jth document
$p(t_i, D)$	The occurrence probability of the ith term in the text D

Part I
Fundamental Concepts and Models

The first part of this book covers the first three chapters and is intended for readers having a linguistic background, or more generally an interest in the humanities. Therefore, the statistical and mathematical notation is kept to the minimum needed, and the linguistic explanation or justification is viewed as more important. The main objective of the first part is to expose to the reader the diversity of problems that can be found under the general heading of stylometry. The observed written style of a document is not fully determined only by the author, but other important factors have key influences, for example, the text genre or the time period. For example, teenagers do not write a tweet like an essay, they do not speak between themselves as they speak with their grandparents. Those simple examples show us the large variability of communication situations and contexts even when analyzing the same author. Moreover, the author himself entails many dimensions that could be studied, such as the stylistic difference according to author's gender, age range, social origin, or psychological personalities.

Of course, the central application of stylometric studies remains the authorship attribution question, with the classic example being whether or not Shakespeare is the true author of his plays. But stylometric methods could also determine some demographic variables about the real author (e.g., Is it a woman? Older than 40 years old? Born in Germany?). The questions and their related concepts are presented and discussed in Chap. 1.

In Chap. 2, the notation and some additional concepts are introduced. Then a set of historical documents (the *Federalist Papers* written in 1787–1788) is presented and will serve as a running example for illustrating the different aspects of each stylometric model or measurement. Moreover, several overall stylometric measures are exposed and examined in order to be able to quantify the main aspects of a document, a given author, or a specific period. Numerical examples are introduced and commented to clearly understand the various steps appearing in the needed computation.

The last chapter of this first part presents in detail four authorship attribution methods usually applied in the humanities or with literary works. Each of these approaches precisely defines a stylistic representation of each text and suggests a

modus operandi to measure the stylistic similarity between pairs of texts or between author's profiles. Finally, they outline a decision rule or a procedure to identify the most probable writer behind a doubtful text or text excerpt. Of course, the presented methods could be applied to solve other questions than authorship attribution, for example, to draw a profile of the author or a stylistic map of a set of documents written in a given period by a group of writers.

Chapter 1
Introduction to Stylistic Models and Applications

During the past 50 years, the volume of data available in electronic format has grown exponentially with the progress accomplished in computer and network technology. In parallel, various statistical tools and methods have been designed, implemented, and are now freely available (e.g., with the R software and its numerous packages). Nowadays, numerous domains of human knowledge view digitized data as a valuable resource and apply machine learning approaches to verify hypotheses or to identify patterns in large datasets. Natural language processing (NLP) and some branches of applied linguistics are following the same direction. Thus the aim of this book is to precisely present and discuss different models and approaches for solving stylometric problems, in particular to solve authorship attribution questions, to discover the author's gender, or to resolve other stylometric questions.

All these questions can be solved by different text classification models, usually based on one sample of examples or instances for which the correct attribution or decision is known, and a second sample (sometimes limited to a single document) for which the attribution is either unknown or doubtful. The main problem is then to understand the most important factors explaining the style differences between authors or more generally between predefined categories such as author's gender, or time periods. Then one can define precisely how a computer system can represent the style of a text or a set of documents, how the similarity between two stylistic representations can be effectively measured, and what degree of certainty can be attached to the proposed attribution. This first chapter provides a broad overview of the different target applications and defines and explains the main concepts of stylometry.

The rest of this chapter is organized as follows. Section 1.1 defines the notion of text classification in the context of stylometric applications. Section 1.2 explains the concept of text style and describes its main explaining factors. Section 1.3 exposes the most well-known problem in stylometry, namely the authorship attribution question and three variants. The next section describes the author profiling question in which some demographics of the author should be inferred from the text

J. Savoy, *Machine Learning Methods for Stylometry*,
https://doi.org/10.1007/978-3-030-53360-1_1

s/he wrote. Section 1.5 illustrates some examples related to forensic linguistics. Section 1.6 exhibits the author clustering problem in which the system must recognize text belonging to the same class (e.g., written by men vs. women, or according to the true author). The last section reveals some additional problems or questions that can be solved, even partially, by considering the stylistic aspects present in a single text or in a sample of texts.

1.1 Overview and Definitions

The main objective of a *text categorization* (TC) or *text classification* system is to automatically assign predefined labels to texts according to their content or style. In this book, a *text* corresponds to a natural language writing not an audio source or a picture of a text (e.g., a scan of a medieval manuscript). The term text or document must be interpreted in a broad sense and can correspond to several forms (e.g., novel, poem, allocution transcript, last will letter, blog post, set of tweets, etc.). The text could be stored in a structured format (e.g., in XML) in which the document structure is clearly marked and the logical components (e.g., chapters, titles, footnotes) can be easily identified. With a semi-structured document (e.g., a web page with its HTML tags), the logical structure is partially provided while an unstructured text (e.g., a transcript of an uttered speech) can be viewed as a stream of words. The document structure could be useful for some applications and is not of prime importance for others. The target text could be limited to a part of an entire work (e.g., a chapter in a novel, a scene in a play, or even a few paragraphs in an e-mail). Associated with a document, one can find tables, graphics, pictures, videos, or hyperlinks. These non-textual elements could be useful in determining the true labels, but the current presentation is focusing mainly on the textual content.

Second, the *labels* represent the possible *categories* of interest. They must be viewed as tags without pertinent and useful meaning for the classification task. These labels could indicate the candidate author names (e.g., Shakespeare, Marlowe, Bacon), the possible text genres (e.g., play, poem, novel), the various keywords or topics, or even a binary answer (e.g., yes or no). When determining the possible labels, several scenarios can be encountered. The set of possible categories could be limited to two (e.g., Is this text written by a man or a woman? Is this document written by a single author?). Usually, the target labels form a set of possible answers and only one must be assigned to the test document (e.g., Is this text written in French, Italian, Spanish, or Portuguese? (language identification)). Sometimes, a few labels can be assigned to a document. For example, in a news agency, an incoming newsflash can receive several keywords such as "technology," "India," and "emerging market."

The set of predefined labels can form a more complex structure such as a tree (e.g., Is this article about sport, business, politics, or culture? And what kind of sport/business/politics/cultural event). In this context, usually more than one label, keyword, or descriptor can be attributed to the input text (e.g., exactly k keywords, or up to a maximum of k). This last problem corresponds to the automatic indexing

process working with a controlled vocabulary (e.g., the Medical Subject Headings (MeSH) contains more than 25,000 descriptors structured into a thesaurus).

Text categorization applications can broadly be subdivided into two principal subdomains. First, the assignment can be performed according to the *semantics* of the document. The main objective is to help the user exploring a large volume of information (e.g., such as in the medical domain with PubMed via the MeSH thesaurus). Filtering is another pertinent tool in which a user can build his profile according to a set of keywords. Incoming documents (e.g., scientific articles, news, e-mails) are then analyzed by the filtering system, and when they correspond to a user profile, the selected texts are sent to the final user. Another classic example is spam filtering, removing non-relevant e-mails in mailboxes.

Second, the categorization process can be based mainly on the *text style*. The content itself is not of prime importance, but the text style forms the core from which pertinent features should be extracted. The style in its broad definition reflects personal choices, usually related to the main intent of the author (e.g., to explain and persuade the reader) but also for aesthetic reasons [395]. The particular style is reflected by the used words, the expressions, or the jargon occurring in the texts. By inspecting the sentence construction, the style includes aspects related to the syntax and grammar.[1] In addition, the repetition rate, the mean sentence length, and the frequent use of particular stylistic figures (e.g., ellipses, similes, metaphors, etc.) are other elements associated with the style. The author choices are not unlimited, and the adopted style is subject to constraints (e.g., oral vs. written communication, text genre, etc.). The style of a remark uttered by G. Washington is clearly distinct from Obama's style, and a colloquial conversation differs from a formal one or from the style that can be observed in a set of tweets.

1.2 Style and Its Explaining Factors

Style can be defined as a manner of expression or way of writing, starting with the choice of words, the combination of two or three words, the punctuation, the sentence structure, the target prosody, the grammar patterns, and all the elements that an author likes to use [301]. For Crystal [73]:

> "By style I mean the set of linguistic features that, taken together, uniquely identify a language user. The notion presupposes that there has been a choice—that someone has opted for Feature P rather than Feature Q (or R, S, ...)."

[1] The grammar specifies how the words are arranged to form sentences while the syntax governs the word order.

The linguistic items defining a particular style can be found at the lexical, syntactical, grammatical, and semantical level, as well as in the text layout. To determine a stylistic element, the notion of *choice* or *freedom* is essential. For example, synonyms offer multiple alternative words or expressions to indicate the same (or similar) concept (e.g., restaurant, coffee shop, bar, saloon, cafeteria, inn, pizzeria, Starbucks, where we met last week, etc.). Koppel et al. [217] suggest exploring this degree of liberty by evaluating a degree of "synonymy." The syntax also offers some freedom to the writer. Of course, for some items the position is fixed, for example, the position of the determiner *the* in the sentence "Now, the cat chases the mouse" (the sequence "mouse the" makes no sense). However, the position of the adverb *now* is not fully fixed and it can appear in another position ("the cat now chases the mouse" or "the cat chases the mouse now"). Furthermore, useful stylistic features are both *frequent* and *ubiquitous*. When writing a sentence and taking account of both the lexicon and the syntax, the author is faced with multiple decisions to achieve the wished target effect.

The stylistic markers do not appear only at the lexical and grammatical stage. At the semantics level, one can analyze the context of some words to define the particular idiosyncrasy of an author. For example, in Corneille's plays, the word *love* is strongly associated with the father's figure (who is usually an obstacle to this love) [228].

More concretely, can we differentiate between the style and the contents of a message? As a good example, Crystal [76] presents this set of sentences (see Table 1.1) reflecting several styles, progressing from a formal style to a casual one.[2]

Table 1.1 Stylistic variations around the same content

The village does not have a post office.
The village has no post office.
The village doesn't have a post office.
The village hasn't got a post office.
The village hasn't got no post office.
The village ain't got no post office.

The style and the contents must not be viewed as fully independent but as two faces of the same coin. The author selects a style to support a message and to reach an objective. However, in this choice several constraints must be taken into account.

As a first and most important explaining factor, the *text genre* explains some lexical or syntactical choices [34, 45]. Writing a sentimental or an adventure novel, an ode, a heroic couplet, a tragedy in verse, or a comedy in prose imposes some limits in the preference for some words or expressions. For example, the

[2] In French, Queneau [312] was able to write the same short story (around two paragraphs) in around 100 different ways.

composition of a poem is governed by a type of rhythm (and sometimes by a fixed number of syllables per verse) restricting the lexical choice of the author. When analyzing stylistic differences between text genres written by the same author, Burrows [45] concludes that different text genres written by the same author present more variability than texts belonging to the same text genre but written by several authors. This result was also confirmed by other studies [16, 30].

To illustrate this, we can mention scientists who prefer using the passive voice. Such constructions allow them to present the facts in a more impersonal and objective form (e.g., "it can be observed . . . ," "spectral analysis was applied . . . "). Comparing texts belonging to distinct text genres, similarities and differences can be discovered and the expression *style of poetry* or *newspaper style* reflects these stylistic resemblances within a given text genre.

The second factor is the *author* himself[3] with his own choice and background (e.g., gender, age range, education, social class, native language, etc.). Individuals have some likes and dislikes about the language and the writing. When facing with synonyms, some persons prefer one word or expression than another, for example, the term *actually* or *in fact* (other examples: while/whilst, because/since, film/movie, etc.). At the grammatical level, some authors produce longer sentences and frequently use the construction *of the*. Other writers prefer using contracted forms when possible. Some authors opt for longer explanations and describe all the details, while others choose concise descriptions. In applied linguistics, all these individual language differences are studied in *stylistics* while the variations between groups separated by social variables (e.g., gender, social position, region) are the object of *sociolinguistics* research.

The *time period* corresponds to the third factor [378] and each period imposes its own stylistic preferences. This aspect is visible in expressions like *classical style*, *postmodern style*, etc. For a more concrete example, one can observe that the sentence length decreases over time. In the eighteenth century, the mean length of speeches uttered by US presidents since 1945. For example, in speeches uttered by Madison under his presidency (1809–1817), the mean sentence length reaches 42 words per sentence while Trump's mean corresponds to 20.5 words [344]. This tendency seems to be reinforced by a fast-paced life (and for some persons by the frequent usage of texting or tweeting).

These first three factors correspond to the most important ones. As additional reasons, the topics have clearly some effect on the lexical choice and more generally on text style. It is known that we can encounter a medical, political, or legal parlance with each domain having its own vocabulary (or *lexis*), idiomatic expressions, and phraseology. When writing a novel, the author may have to provide the correct terms to describe a harbor or the words occurring in a dialogue between two sailors.

[3]To simplify the presentation, the masculine form has been selected to indicate equivalently a man or a woman.

The *communication type* also plays a role in the style. We do not write as we speak and when using web-based communication channels, we can adopt new features (e.g., less strict orthography, emojis) [74, 256]. In a tweet, one can write "C U" to say *see you* but the former will never appear in a newspaper article. In oral, the speaker pronounces more pronouns compared to a written text, repeats the same expressions more frequently, and thus presents a less abundant vocabulary [75].

Lastly, the *audience* has an influence on the style (e.g., look at the language differences between an official speech or a colloquial discussion) [187]. In an informal discussion, one can use *gonna* but will opt for *going to* in another context. One could also mention the editor as a possible source of stylistic variations, and its impact seems to be related to the text genre [326]. The punctuation can be the object of this variability because, for example, the usage of the comma is controlled by imprecise rules that can be fixed and imposed by the editor [78].

When taking into account all of these stylistic factors and their implications, we reach the conclusion that defining a Shakespeare's style reflecting his entire work is impossible. As for all authors (e.g., Goethe, Dante, Proust, etc.), writing a poem for a given audience, a play in prose or verse, a tragedy or a comedy, and in a specific time period and context imposes constraints on the word choice, syntax, or grammatical constructions. Thus, one cannot speak about a unique Shakespeare's style, but each author presents different stylistic facets.

Finally, three clarifications must be provided. First, the term *author* covers different aspects. In this book, the author is the person who composed the text with his corresponding lexical and syntactical choices. It is not the person who writes the selected words on a paper or using a word processor (the copyist or amanuensis). Likewise, the author is not the person who elaborates the scenario, the thoughts to be expressed (e.g., in a testimony), or the different characters or figures appearing in a novel or a play.

Second, the term *traditional authorship* must be interpreted as manual investigation to determine the true writer [246]. Our focus is on computational and statistical-based methods to solve various text categorization questions based on the text itself. In this case, the term *stylometry* covers this scientific activity. Unlike traditional authorship investigation, external proofs such as the historical context of the produced work, bibliographical evidence, or physical analyses (handwriting, watermarks, chemical analyses of the ink or paper) are ignored.

Third, even if our examples are given in English, the described methods and practices tend to work well for all natural languages, based on letters, syllabaries (e.g., Hangul (Korea), Katakana (Japan)), or sinograms (Chinese). To the best of our knowledge, we are not aware of studies demonstrating a real divergence in the effectiveness of methods resulting solely on the script difference between two languages.

1.3 Authorship Attribution

Authorship attribution or author identification [163, 183, 190, 219, 246, 278, 325, 373] is a well-known problem in stylistics and certainly the most studied question in stylometry. This problem can be stated as follows. Given a set of texts with known authorship, can we determine the author of a new unseen document? Under this general definition, three distinct contexts can be encountered.

First, the *closed-set* attribution problem assumes that the real author is one of the specified candidates from whom a sample of texts is available. This is the typical problem that can be found in research papers. Usually, the list of possible candidates has been determined from different external evidences (e.g., selecting because they are in the same text genre and presenting similar styles as the disputed document). In many cases, such a list contains a few names. For example, several studies suggest that behind Shakespeare's (1564–1616) signature one can discover another name. In this authorship debate, the most cited possible true authors are Sir F. Bacon (1561–1626) (the favorite candidate during the nineteenth century and the beginning of the twentieth century), E. de Vere (17th Earl of Oxford, 1550–1604), C. Marlowe (1564–1593), W. Staley (6th Earl of Derby, 1561–1642), or J. Florio (1553–1625) [104, 262, 388]. For others, one must not look for a single person behind Shakespeare's works but a group of authors. However, knowing that Shakespeare lived from 1564 to 1616, the candidacy of Marlowe or that of de Vere presents some temporal overlap issues. A similar question appears in French literature with the hypothesis that the most famous plays known to be written by Molière (1622–1673) were in fact written by P. Corneille (1606–1684) [227].

If these questions concern a set of works published in the Early Modern English era, other authorship issues focus on parts of a particular work. Knowing that several plays appear with two names (e.g., the play *The Two Noble Kinsmen* was written jointly by J. Fletcher and W. Shakespeare), the authorship question is to determine which scenes have been written by the first and the second author [65, 415]. As another example, it is admitted that the play *Henry VI* was, in part, a joint work between Shakespeare and Marlowe (mainly in Parts 1 and 2).

Second, in the *open-set* context, the real author could be one of the proposed authors or another unknown one. This case requires a more complex attribution method allowing a *don't know* answer due to lack of evidence, insufficiency, or failure of proof for all possible candidates. This problem instance is less studied in research papers, certainly because most of the attribution methods tend to provide an answer even when the stylistic evidence is rather weak and not fully convincing.

Third, the *verification* question provides a binary response as to whether a given author did in fact write a given text [214, 215, 218]. As input, the system has a sample of texts written by the unique candidate on the one hand, and on the other, the test document. This problem should be viewed as the more general one [221]. As soon as you can provide an automatic and effective model to solve this question, you can solve the two previous ones. In fact, having for each possible author a sample

of their writings, the two previous questions can be transformed into a sequence of verifications, one per candidate author.

One of the oldest authorship attribution questions is precisely a verification one: determining whether or not St Paul is the real author of the *Epistle to the Hebrews*. De Morgan [86] suggests using the distribution of word lengths (the number of letters per word) as a stylistic indicator to solve this problem. As a more recent problem, one can ask if McCartney is the single author of the song *In My Life* [131].

To solve these three distinct authorship questions, one important and often hidden hypothesis is the stability of each individual style denoted as stylistic fingerprint or stylistic idiosyncrasy. "Le style c'est l'homme" (the style is the man) said Comte de Buffon, suggesting both a stylistic stability and a way to discriminate between several authors. It is assumed that, for a mature person, his stylistic markers and language patterns will not undergo a large change during his life [224]. Such a *stable* stylistic identity must be interpreted as the style of an adult. For Joos [187], the language and style correspond to five periods in a human life, namely the baby language style, then the child, teenager, adult, and finally the elder one. Each of these frames presents its own language patterns and style (both in the lexical and syntactic constructions), but the mature phase is certainly the longest. Thus, it is a good practice not to compare texts written when the author was a teenager to those produced later. Moreover, when comparing two works, the time span must be taken into account when the publication date varies more than two decades. For example, Hoover [169] found that two novels written by the same author but with a difference in the publication date of more than 30 years are difficult to be assigned to the same writer (see also [50, 121, 171]).

As output, the system can provide a single name (or a binary answer in case of verification). However, it is more common to return a ranked list of names with a score indicating the belief or degree of certainty that the corresponding author is the true one. A clear and precise interpretation of such values is rather difficult for the user. Thus, within some methods, a probability can be estimated for each possible answer leading to a better information and clearer interpretation of the result. Finally, the proposed attribution should also be supported by a linguistic reasoning or by highlighting some language pattern similarities that cannot occur simply by chance.

1.4 Author Profiling

In some issues, the true name of the author is not of prime importance. The focus is on author profiling [317] or authorship characterization with the objective to identify some demographics about the writer such as his gender, age range, native language, social status, or even some of his psychological traits [21, 40, 280].

From those variables, the gender distinction might be viewed as the simplest one. The classification decision is binary and a relatively large amount of textual data can be collected. However, such a classification system can be effective only if

the writing style between genders does differ [97, 428] on the one hand, and on the other, if such differences can automatically be detected [297]. In addition, it must be recognized that there is a continuum in style between an extreme male and female figure. Moreover, it is not clear whether or not LGBT people can show other distinct writing variabilities compared to prototypical male and female figures.

Due to a large disparity during the language acquisition period, it is hard to have a good estimate of the age range for babies or children [299]. Therefore, the first age range considered by a profiling system is usually teenager (e.g., from 16 to 20 years old). As a simple binary discriminative model, two categories can thus be considered such as teenagers vs. older persons (e.g., 25 and more). In another perspective, four to five age ranges can be created (e.g., 18–24, 25–34, 35–49, 50–64, and 65+) assuming that stylistic markers can be detected to differentiate between these age ranges. Usually, however, the text contents reflect aspects that can be useful to discriminate between some age ranges (e.g., *marriage* is more related to the 25–34 age range, while *mortgage* or *children* could be associated with the 35–49 age range). For example, for writers born before 1950, the masculine form is employed to denote both genders while younger authors prefer to use more frequently the expression *he or she* or *s/he* (a third-person neutral pronoun does not exist in English). In addition, younger writers tend to depict a larger stylistic variability [291] and usually present more feminine stylistic patterns than older ones [299].

In Tables 1.2 and 1.3, two short blog posts illustrate the text style difference between a male and a female writer, as well as the author age range differentiation. Usually, readers do not have difficulty in identifying a teenaged woman as the author of the passage shown in Table 1.2. As explanation, one can argue the presence of emotions (e.g., ashamed, cry) for determining the author's gender. For defining the age range, a good and precise explanation is harder to put forward. One can argue about a jazz competition and that the words appear simple, belonging to a basic vocabulary, leading to a teenager behind the text depicted in Table 1.2.

Table 1.2 A first example of a blog post. Written by a man or a woman and which age range?

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotten, and I wanted to cry, but... it's ok.

With the example depicted in Table 1.3, the identification of the author's gender is more problematic but the age range seems easier (this is written by a male writer, between 25 and 35 years old). But knowing that pronouns are used more frequently by female writers and that nouns appear more often with males simplifies this identification task. Just counting the number of {*I*, *me*, *he*, *we*} on the one hand, and on the other the number of nouns (or the number of {*the*, *of*, *in*, *this*}), one can detect the author's gender. Of course, other features can discriminate between

Table 1.3 Another example of a blog post. Written by a man or a woman and which age range?

My gracious boss had agreed to let me have one week off of "work." He did finally give me my report back after eight freakin' days! Now I only have the rest of this week and then one full week after my vacation to finish this damned thing.

a male and female such as a higher frequency of emotion words and negations for women, or swear expressions for men.

The time period has also a clear impact on the style. To illustrate this, read the text depicted in Table 1.4, a passage of a political declaration from the eighteenth century (written on July 4th, 1776, by T. Jefferson[4]). Nobody speaks like this today. First, this excerpt corresponds to a single sentence, a very long one with 71 words. Second, many words are capitalized (e.g., "Course," "Nature," "Law"). Third, the tone is very formal and solemn providing reasons justifying an important decision.

Table 1.4 Excerpt from the US declaration of independence

When in the Course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation.

The language used can also reveal the geographical origin both at a national or international level. Obviously, the different accents play the most important role in identifying where a speaker comes from [79]. In a message, the vocabulary, the grammar, and the semantics might vary from one region to another, variations studied in dialectology. For example, the verb *to be*, the negation, and the different contraction forms show variability across England, from the *it's not* (in the North) to *ain't* (East Midland or South). The word order can also switch from *give it to me* (South-West) [71].[5] At the lexical level, the same object could have different denominations (e.g., "on vacation" in the United States, "on holidays" in UK, shop vs. store, post vs. mail, or line (railway) vs. track).[6] The spelling is also affected with the well-known differences between US and UK English (center vs. centre, color vs. colour, etc.) [84]. The grammar also diverges from one country to another

[4]T. Jefferson wrote the first draft of the independence declaration. After this document was edited by a committee of five members, and then by the whole Congress.

[5]All these language differences are a matter of critical judgments and could lead to a prestige scale and mockery (which have absolutely no linguistics support).

[6]Such differences appear in other languages such as in French with the term *mobile phone* called *cellulaire* in Montreal, *natel* in Geneva, or *portable* in Paris.

(e.g., with some Caribbean features as in "They going with two car," missing both the verb *to be* and the plural suffix). At the semantics level, the same word could cover different meanings, for example, the term *robot* meaning traffic light in South Africa. These language variations could differ between the official one and the usage. For example, even if Canada has adopted the UK standard English, in practice the US dialect is used.

Other aspects of the author could have an influence on the resulting style. The social position can be reflected by the choice of some words instead of others (*sick* vs. *ill*). Persons having a higher social status will employ a richer vocabulary and use more articles, prepositions, and longer words [75, 299]. The social group also has an influence on our language. In everyday language, one meets the expressions *journalist style* [19] (e.g., factual, and appealing to a wide audience) or *politician tone* [154, 155] (e.g., to achieve an angry/confident mood from the audience).

The native language of a bilingual writer can be discovered in a text. For example, a native French speaker tends to use words coming from a Latin stem more often instead of a Germanic synonym (e.g., adorable vs. lovely). In addition, this person would add a space before the question mark or the semi-colon punctuation symbol.

Finally, some psychological traits, ways of thinking, or the current mood of the author can be detected by inspecting his writings [280]. Sometimes one can call this the voice (or personality) of the author. For example, Pennebaker [299] indicates that angry people employ negative emotions more often as well as second-person (you) or third-person pronouns (he/she/it/they). They prefer the present tense. If sadness can be detected with these same words, sad persons tend to employ the past or future tense more than the present one. In a recent study, Noecker et al. [283] show that some psychological variables (judgment, feeling, extraversion) can be identified with a relatively high accuracy (around 85%) but others are clearly more difficult to predict (way of thinking, perception with an effectiveness around 60%). In a similar analysis, a person having depression or anorexia will depict some specific language patterns (e.g., a higher usage of *I/me/my*) and negative emotional words or anger expressions (e.g., *sad*, *cry*, *pain*), together with a decrease in the frequency of third-person pronouns [61, 143, 244, 245].

1.5 Forensic Issues

Identifying the author of a text can be useful to resolve offenses or to establish a criminal profile of a perpetrator (*forensic linguistics*). In this context, features present in the speech (voice printing analysis) or detected in a handwriting analysis can be pertinent to identifying the individual or to providing some demographics about the possible author. Even limited to writings, the cautious analysis of offender texts (e.g., ransom letter, threat e-mail) can reveal a few linguistic patterns to support police investigations [289, 290].

For example and as described in [291], the author of a threat message was identified by his systematic incorrect spelling of some terms or formulations (e.g., *alot*, *aswell* and the recurrent use of *stuff* instead of *staff*) and by writing "?!" instead of a single question mark. In another case, the occurrence of rare or very unusual words or expressions can establish evidence to discover the possible author (e.g., *covfefe*[7] by Trump (May, 2017) or *lodestar* used by an anonymous opponent layperson at the White House (Sept., 2018)).

In this kind of investigation, the main concern is the text length, usually short, for example, a few SMSs or tweets. In such circumstances, the application of identification methods is less predictive and the error rate could be too high to be admitted in a court of law [54, 64]. The stylometry analysis could however be useful for police officers, for example, by reducing the number of possible suspects (for an example, see [137]).

Other issues could be solved by considering the strong similarities between two texts or two passages such as plagiarism detection [10, 27, 379]. As a rule of thumb, to discover such an awful practice, one can consider that an identical sequence of five words between two passages is a strong indication that both are coming from the same source or than the second is a copy of the first one [291]. Even if the occurrence probability of an identical sequence of five words is not zero, it is very small, apart from some named entity denominations (e.g., former vice president of the United States of America, chief executive of the Royal Bank of Canada (in short, CEO of RBC)).

Other forms of plagiarism are more difficult to detect by a simple comparison. Instead of a simple copy/paste operation, the source could be text paraphrased or rewritten with a set of possible synonyms and changes in the sentence construction. For example, in his inaugural speech (1961), Kennedy said, "Ask not what your country can do for you, ask rather what you can do for your country," but a similar sentence was uttered by President Harding in 1923 "We need to be thinking not so much of what the country can do for us but what we can do for our country" [176].

An automatic detection tool needs therefore to operate with a soft matching algorithm. The aim is then to match similar sentence construction and meaning but written with some lexical variations. In addition, the source message could be written in one language and the plagiarism in another, usually by simple machine translation [307]. With numerous text repositories, search engines, and freely available machine translation tools, the Web has facilitated all these activities. As a corollary, the demand for automatic detection systems of such practices has also increased.

[7] According to *The Independent*, *covfefe* means *coverage* and the full tweet was "Despite the constant negative press covfefe." The White House Press Secretary Sean Spicer confirms it was not a typo, but "the president and a small group of people know exactly what he meant" [110].