Journal of Universal Computer Science, vol. 26, no. 1 (2020), 156-170
submitted: 30/12/18, accepted: 15/11/19, appeared: 28/1/20 CC BY-ND 4.0

Authorship Studies and the Dark Side of Social Media Analytics

Patrick Juola
(Evaluating Variations in Language Laboratory, Duquesne University
Pittsburgh, USA
juola@mathcs.duq.edu)
(Juola & Associates
pjuola@juolaassociates.com)

Abstract: The computational analysis of documents to learn about their authorship (also known as authorship attribution and/or authorship profiling) is an increasingly important area of research and application of technology. This paper discusses the technology, focusing on its application to social media in a variety of disciplines. It includes a brief survey of the history as well as three tutorial case studies, and discusses several significant applications and societal benefits that authorship analysis has brought about. It further argues, though, that while the benefits of this technology have been great, it has created serious risks to society that have not been sufficiently considered, addressed, or mitigated.

Key Words: authorship attribution, social media, computer security
Category: K.4.2; K.6.5; I.7.m; I.5.4

1 Introduction

Scholarly disagreements about authorship are nothing new. The anti-Stratfordian controversy about Shakespeare's canon is well-known [Friedman and Friedman, 1957], even if considered a fringe theory. Perhaps less well known (at least in Anglophone scholarship) is the question of whether Molière wrote his work,[1] or whether it was written by Pierre Corneille, perhaps as part of a "wide system of ghostwriting" [Cafiero and Camps, 2019] in 17th century French Theater.

Although this type of scholarly controversy goes back millennia, and attempts to resolve this sort of question on the basis of empirical analysis and data date to at least the 19th century [de Morgan, 1882, Mendenhall, 1887], recent advances in artificial intelligence, machine learning, and natural language processing have made this type of analysis much easier, faster, and more reliable. Applications of authorship analysis include not only the literary [Collins, 2013, Cafiero and Camps, 2019], but also the historic [Mosteller and Wallace, 1963, Mosteller and Wallace, 1964], the journalistic [Brooks and Flyn, 2013, Brooks, 2013, Herper, 2014, Pesca, 2018], and even the legal [Leonard, 2006, McMenamin, 2011,

[1] See https://en.wikipedia.org/wiki/Molière_authorship_question; accessed 18 December 2019

Ainsworth and Juola, 2019]. This technology has proven to be a useful scholarly tool across the academy.

At a broader level, though, authorship analysis has created new and serious threats both to individuals and potentially to society as a whole. In combination with the popularity and omnipresence of social media, which provides a fertile ground for individual analysis, this technology has a high risk for abuse. This paper discusses authorship studies, the underlying technology, some applications, and some nightmare scenarios about how these applications can be misused and the damage they can cause.

2 Authorship Attribution

Authorship attribution [Juola, 2006,Koppel et al., 2009,Stamatatos, 2009,Grieve, 2005, Jockers and Witten, 2010], strictly defined, is the task of analyzing a document to determine, not its meaning, but the identity of the person who created it. For example, a recent paper [Cafiero and Camps, 2019] addressed the question of Molière's writings with a large-scale study of various attributes of his work, including lexicon, rhyme, morphology (including a separate study of affixes), morphosyntax, and function words, comparing writings traditionally attributed to Molière with some of his contemporaries. They found that Molière's writings displayed both the internal stylistic consistency as well as the systematic differences from others that would be expected if he had been writing his own work.

So how does this work? The theory of authorship attribution is well-expressed by Coulthard [Coulthard, 2013]:

> The underlying linguistic theory is that all speaker/writers of a given language have their own personal form of that language, technically labeled an idiolect. A speaker/writer's idiolect will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in the written and spoken communications they produce. For example, in the case of vocabulary, every speaker/writer has a very large learned and stored set of words built up over many years. Such sets may differ slightly or considerably from the word sets that all other speaker/writers have similarly built up, in terms both of stored individual items in their passive vocabulary and, more importantly, in terms of their preferences for selecting and then combining these individual items in the production of texts.

Some of these differences are obvious and easily explainable. A person who speaks of "a lorry," who buys "paracetamol" at "a chemist's shop" and who spells "colour" with a 'u' is presumably the product of a non-US education system.

Others may require some unpacking; for example, an early legal case [Shuy, 1998, Leonard, 2006] involved a kidnapping and a ransom note that read, in part, "*Put [the ransom money] in the green trash kan [sic] on the devil strip at the corner 18th and Carlson*." The term "devil strip" will probably not be familiar to most readers. It is a term used for a strip of grass between the street and the sidewalk, but it is only used in the Akron, Ohio area. As such, the examining linguist was able to pinpoint where the author of the ransom note was from, and, based on this information, the police had no trouble solving the case.

However, some differences do not admit of easy explanation. One of the most powerful and reliable methods for determining authorship is the use of so-called "function words." These are the small, frequent words like (in English) "the," "of," "and," and so forth. Most of them are closed-class words such as prepositions, conjunctions, articles, and pronouns. More importantly, they tend to carry little content or meaning of their own (and hence are omitted from most natural language processing tasks), instead describing relations among other words in the discourse. Most importantly, because they have so little meaning of their own, they are typically found in all types of documents by all sorts of authors. As expressed in [Cafiero and Camps, 2019][p. 2–3]:

> According to the literature of the past 3 decades, the analysis of function words is the most reliable method for literary authorship attribution. The underlying intuition is that function words are used mostly according to unconscious patterns and vary less according to the topics and genre of the texts. Psycholinguistics studies have shown that function words are perceived by the readers on a less conscious level and are read faster than content words; they might also be chosen less consciously by the writers, while nonetheless being able to convey significant information on the speaker or writer.

As an example of the use of function words, consider the task of describing where the fork is in a standard table setting. From left to right, the setting typically includes fork–plate–knife–spoon. However, there are many possible descriptions for the location of the fork. It could be

- *to* the left of the plate (https://en.wikipedia.org/wiki/Table_setting)
- *on* the left of the plate (https://www.etiquettescholar.com/dining_etiquette/table_setting.html)
- *at* the left of the plate (https://www.janinestone.com/luxury-lifestyle/porcelain-designer-teaches-table-setting-etiquette/)
- *at* the left *side* of the plate (Pacific Rural Press, September 6, 1913, p. 236)

among many others. There does not appear to be any sociolinguistic explanation (such as age, gender, or region) for this variation, nor is the meaning altered. Indeed, psycholinguistic research [Bransford et al., 1972] suggests that people may not even notice or be able to remember which version was used. However, these differences are easily detectable and robustly measurable, which makes them useful for determining authorship.

3 Some case studies

Binongo [Binongo, 2003] provides an easily understandable example of how this works in his study of the fifteenth book in the *Oz* series, *The Royal Book of Oz*. Authorship of this book has been disputed between L. Frank Baum, the original author of the series (including the first book, *The Wonderful Wizard of Oz*) and Ruth Plumly Thompson, the woman hired to continue the series. The *Royal Book* was published in 1921 with Baum's name as author, "enlarged and edited" by Thompson, but some have suggested that the 15th book was almost entirely Thompson's work.

As is typical in this type of problem, Binongo treated this as a standard text classification problem. He collected two training sets, one of Binongo's undisputed *Oz* novels, and one of Thompson's. The test set was, of course, the *Royal Book* itself. He then selected the fifty most frequent function words as a feature set. He broke the texts into 5000 word blocks and calculated the frequency of each word in each block. Finally, he applied principal component analysis (PCA) to produce a two-dimensional approximation of the original fifty-dimensional space.

This analysis clearly showed that Baum's and Thompson's writing styles differed. In the final image, all undisputed Baum samples were to the right of the origin (indicating a positive value on the first and most significant principal component), while all of Thompson's samples were on the left. Furthermore, analysis of additional samples from non-Oz works showed the same left-right separation. Finally, analysis of the *Royal Book* showed all samples to be on the left side of the origin, in line with Thompson's style but not Baum's. Binongo concluded that the 15th book "was written in Thomson's Pen."

Another well-known example is the study of the *Federalist Papers* by Mosteller and Wallace [Mosteller and Wallace, 1963, Mosteller and Wallace, 1964]. While the authorship of most of the papers is straightforward, twelve papers were claimed by both Alexander Hamilton and James Madison. Again, this is a simple classification problem; the undisputed Hamilton papers form one training set, the undisputed Madison papers the other, and the disputed works become the test set.

Mosteller and Wallace [Mosteller and Wallace, 1963, p. 276]observed that, while, in some stylistic regards

> [h]high rates for 'by' usually favor Madison, low favor Hamilton; for 'to' the reverse holds. Low rates for 'from' tell little, but high rates favor Madison." Hamilton used the word "while," but Madison used "whilst."

Mosteller and Wallace stated that "the best single discriminator we have ever discovered is 'upon,' whose rate is about 3 [instances per thousand [words] for Hamilton and about 1/6 per thousand for Madison."

They applied the then-novel technique of naïve Bayes classification (NB) to the observed distributions of thirty hand-selected words and concluded that all of the disputed papers had been written by Madison. This conclusion has been strongly supported by later scholarship [Martindale and McKenzie, 1995, Rockeach et al., 1970, Holmes and Forsyth, 1995, Rudman, 2005, Tweedie et al., 1996, Jockers and Witten, 2010].

One of the best known examples of a real-life controversy addressed by authorship attribution is Juola's analysis of *The Cuckoo's Calling*. [Brooks and Flyn, 2013, Brooks, 2013, Juola, 2013a, Juola, 2015] Published in 2013 under a pseudonym, an anonymous Twitter user suggested that the true author was J.K. Rowling, of *Harry Potter* fame. At the request of a British newspaper, Juola collected a corpus of other Rowling novels as well as works by three distractor authors. Using a variety of different feature sets, including word length, use of (the 100 most frequent) function words, overall vocabulary (all words), and word 2-grams (pairs of adjacent words, which provide insight into syntax as well as lexicon), he applied a nearest neighbor (k-NN) classification algorithm to determine which author was most similar to the author of *Cuckoo* along each of the four feature sets. Juola concluded not only that Rowling was the most likely candidate among the studied authors, bu, given how consistently she was the most likely candidate, it was highly unlikely for the correct answer to be none-of-the-above. He therefore suggested to the newspaper that, yes, the real author was Rowling. Rowling herself acknowledged authorship the next day, providing a nice validation of this technology.

4 The technological underpinnings

As a classification and machine learning problem, the most important technical questions are "what are the relevant features of the objects of study?" and "what method of learning and classification is to be used?" All three of the studies in the previous section focused on function words as features, but it may or may not be significant that they used slightly different sets of words. Juola also used additional features in analysis not considered by the other two. Other researchers have proposed—and used—many other feature sets. The idea of using word length [de Morgan, 1882, Mendenhall, 1887] as a marker of authorship predates Mosteller and Wallace by a century. Other researchers [Juola, 2006] have proposed using sentence length, vocabulary complexity, vocabulary overlap, synonym pairs, parts of speech, and many others. One of the most powerful and commonly used feature sets are character clusters (formally called N-grams) [Cavnar and Trenkle, 1994, Stamatatos, 2013, Mikros and Perifanos, 2013, Cafiero and Camps, 2019], groups of N adjacent characters without regard to word boundaries—a n-gram can be at the beginning, end, or the middle of a word, or even incorporate the end of one word, a separating space, and the beginning of another. Rudman [Rudman, 1998] has estimated that more than a thousand feature sets have been used successfully in authorship attribution studies. In light of this, the differences in function word selection appear to be more a sign of robustness than of cherry-picking.

By contrast, all three authors in the section above used different classification: Binongo used PCA, Mosteller/Wallace used NB, and Juola used k-NN. Other researchers have used linear discriminant analysis (LDA) [Chaski, 2005, Baayen et al., 2002, van Halteren et al., 2005], support vector machines (SVM) [Diederich et al., 2003, Argamon and Levitan, 2005, Sousa-Silva et al., 2011], decision trees/random forests [Zhao and Zobel, 2005, Khonji et al., 2015], artificial neural networks [Matthews and Merriam, 1993, Merriam and Matthews, 1994, Tweedie et al., 1996], and deep learning [Gómez-Adorno et al., 2018] (among others). Large scale testing [Vescovi, 2011, Juola, 2012] has failed to find any consistent "magic bullet" that significantly outperforms other methods in accuracy. Ensemble methods using multiple datasets and classifiers are practical [Juola, 2008, Juola, 2015, Cafiero and Camps, 2019] and are becoming standard practice as a way of maximizing the accuracy and reliability of authorship judgments.

These techniques have also been shown to be robust to different languages. While most of the cases described above involved works in English (the Molière documents were in French), other researchers have studied Arabic, Spanish, German, Greek, Turkish, Chinese, Japanese, and even indigenous languages such as Arapaho [Juola, 2018] and Kinyarwanda [Illibagiza Umulisa, 2019]. The technology has been proven to work on documents in historical dialects such as Old Church Slavonic and Middle English [Juola, 2006]. There are even results suggesting that authorship attribution can be performed across languages (for example, training documents in English, test documents in Greek) [Juola and Mikros, 2016b, Juola and Mikros, 2016a, Juola and Mikros, 2017] based on cognitive universals of the writer's mind/brain. This is an active research area and will improve over time as a matter of course.

So if we can, in fact, determine authorship, of what practical use is it? The examples given above are primarily literary, although the Rowling case involved a public dispute covered in the newspapers, making it arguably a journalistic application. Other journalists have covered cases like the inventor of Bitcoin [Herper, 2014, Cohen, 2014] or the "Resistance" op-ed in the New York Times

[Pesca, 2018]. It is also of interest to educators, for example, to identify plagiarists [Juola, 2017], especially in the case of (ghost)written-to-order papers that would not be available to search engines such as Google or to plagiarism checkers such as Turnitin. As alluded to above, law enforcement and lawyers are also interested in the possibility of identifying the author(s) of documents. Chaski [Chaski, 2005, Chaski, 2007] describes the real-life case of an ostensible suicide note typed on a shared computer, and the possibility of the note being typed by the murderer instead. McMenamin [McMenamin, 2011] testified about a single possibly-forged email that could shift the ownership of a multibillion dollar company. Juola [Juola, 2013b] describes a case of a political activist seeking asylum on the basis of his anonymous criticisms on the Internet of a foreign government. Ainsworth and Juola [Ainsworth and Juola, 2019] describe 13 legal applications based on real-life disputes. It is clear, then, that the development of this technology can have substantial positive impact both on scholarship and on society.

5 Authorship profiling

In addition to the inference of identity, the same technology can be used to "profile" authors [Argamon et al., 2009], inferring, not their identify, but other attributes. For example, the existence of gender-related differences in language is well-known. At least some of these differences can be detected in text [Koppel et al., 2002, Koppel et al., 2003, Corney et al., 2002, Kucukyilmaz et al., 2006, Hota et al., 2006]. The same classification paradigm as in authorship attribution can be applied: collect training sets of male and female writings, select appropriate features and classification algorithms, learn the differences between the two sets, and classify novel documents appropriately. Argamon et al. [Argamon et al., 2009, Argamon et al., 2005] applied this to classify blogs by gender, by age, by native language, and by personality (using the "Big Five" personality taxonomy). Other researchers have been able to determine traits like education level, social class, country of origin, and many more.

These results have been replicated by many other researchers. Luyckx and Daelemans [Luyckx and Daelemans, 2008b, Luyckx and Daelemans, 2008a] and, independently, Noecker Jr et al. [Noecker Jr et al., 2013], were similarly able to profile people using the better-known Myers-Briggs Type Indicator (MBTI) personality taxonomy. Using keystroke data collected in a simulated office environment, Juola et al. [Juola et al., 2013] were able to profile for, among other things, MBTI personality, gender, and even dominant hand. They were also able to infer self-esteem [Juola and Noecker Jr., 2014].

Other studies have been able to accurately identify bipolar disorder [Noecker Jr. and Juola, 2014, Sekulić et al., 2018], depression [Havigerová et al., 2019], and Alzheimer's [Kernot et al., 2017]. Benton et al. [Benton et al., 2017] discuss many other mental profiling tasks, including the detection of suicidal ideation, schizophrenia, post-traumatic stress disorder (PTSD), anxiety, eating disorders, panic attacks, or simply neuroatypicality (atypical mental health) generally defined. Using data collected from Twitter, they were able to detect these mental states at rates substantially above chance. While the results they obtained were probably not accurate enough to be used as diagnostic tools, this (again) is an area of active research and results will only get better.

6 The dark side

Unfortunately, the affordances of this particular technology easily lend themselves to serious abuses. While the ability to infer mental states noninvaisively may provide a tremendous boost in the ability of mental health professionals to provide care remotely, that same noninvasiveness makes it possible, even practical, to perform a stealth assessment of a person without their knowledge or consent.

As of this writing (2019), it's a fairly standard practice for potential employers to check social media profiles of job applicants as a routine hiring practice. According to CNBC [O'Brien, 2018], 70% of all employers make such checks, looking for red flags such as inappropriate content, signs of drinking or drug abuse, and/or discriminatory content. Social media is similarly used to evaluate candidates' professionalism (for example, did they bad-mouth previous employers or co-workers?) and communications skills. Even an unprofessional screen name can be grounds to eliminate a job candidate at more than 1/5 of hiring companies.

It's a small step from looking at social media by hand to using content analysis algorithms. These algorithms have advantages in speed, in cost, in consistency, and in objectivity. But the same computers that analyze the content can easily analyze and profile the attributes of the author, for example, by looking for signs of "neuroatypicality" or more specific attributes that the would-be employer considers undesirable. From a purely economic perspective, it may make sense as a way of proactively reducing medical costs and reducing medical-related absenteeism. Whether or not it violates the legal rights of the applicant will, of course, vary with the jurisdiction. However, performing a pseudo-medical test on the applicant *without his consent or even knowledge* is a serious violation of international standard on medical care and of the applicant's human rights as defined by, for example, the WHO.

Unfortunately, there are no obvious ways to prevent this type of gross abuse. While anti-discrimination laws could be extended to cover this type of analysis, the fact that a company is using such software as part of its (confidential) hiring process, and the specific details of the software that it is using, are unlikely to be well-known or understood. The ability of applicants to challenge companies under the new law is likely to be limited.

To see how limited this ability is likely to be, look at the risk assessment tools used by the justice system, for example, to determine if a person will be released on bail, to set the sentence, to pick suitable conditions of their sentence, and/or to decide whether they will receive probation and/or parole. Osoba and Welser [Osoba and Welser IV, 2017] provides an excellent survey of some of the biases in this type of system and some of the societal problems it exacerbates. For example, a 2016 report [Angwin et al., 2016] found that a widely used program was 77% more likely to predict a black defendant as being at a high risk of committing a violent crime in the future (and 45% more likely to commit any sort of crime), even when controlling age, gender, and history. Racial discrimination is baked into this software. However, this type of biased analysis can be enough to persuade judges to overturn plea deals and to impose much harsher sentences, and affected defendants have had little success in challenging these decisions.Would a hiring manager be similarly persuaded by a computer-generated analysis indicating that a particular candidate would be a bad fit for a job?

A further risk comes from the ability of authorship analysis to link documents with common authorship. A standard recommendation for personal safety and security on the Internet is not to use your real name or identifying information. If you need to use your real name for professional reasons, keep it separate from your personal activities. However, this precaution is of limited use if a computer can look at the writing style of a pseudonymous Reddit poster and link it to a named professional web site. This not only exposes further information about the user that they might have preferred to keep private, but also provides much more information for the algorithms to mine, further increasing their own capacity.

A potentially worse possibility is the ability of bad actors to use an algorithmic profiling capacity to target specific people. Companies are very interested in the demographics of people who like and dislike their products; the ability to look at comments on Amazon and use that to figure out whom to target is invaluable. But what other capacities does authorship profiling offer?

Cambridge Analytica is, or was, a firm specializing in "psychographic targeting," targeting of ads to people based on their psychological attributes. The primary basis of their technology was the analysis of Facebook "likes" [Bachrach et al., 2012, Markovikj et al., 2013] and other data (Twitter feeds, browsing histories, phone-call patterns) [Gibney, 2018] to determine attributes of individual people, then to target those individual people with specific messages personalized to their personality traits. According to a CA insider and whistleblower, [Scott, 2018] the company used its data to help the "Leave" campaign in the 2016 British EU referendum as well as the 2016 Trump campaign in the United States. As revealed in 2018, a major source of their information was improperly obtained through a combination of lax security on Facebook along with with a deceptive app that handed over not only the user's personal information, but (unknowingly) that of their Facebook friends as well. This security breach as well as the scale of the operations helped propel the CA scandal into international news.

The overall effectiveness of Cambridge Analytica in swaying votes is disputed, but still worrisome. Whether or not CA could persuade people to vote a particular way depends largely upon their specific capacities in 2016. A later company, with better tools, is likely to be more effective at manipulating people based on their individual psychological triggers. At what point does this kind of automated personal manipulation in electoral outcomes become a serious problem?

Consider, though, that the primary violation was not CA's attempts to influence elections, but the deceptive methods by which they gathered their information, information that the users considered to be private. Facebook, for example, had no issue with the idea of psychographic advertisement, but suspended CA over what it considered to be a violation of the terms of service in its data gathering. It is possible, even likely, that similar profiles could be gathered from other information that users have themselves *published*, or in other words, specifically indent to be read by the public at large. A rational user recognizes that anything posted to Twitter, to public discussion forums such as Reddit, or even to a vendor's comment section can be read by anyone. Indeed, that's often the entire point of such postings. Would users be as calm about posting publicly on the Internet if they knew the amount of information about themselves they were also revealing?

At the extreme, this type of analysis can expose participants to personal risk. Using authorship profiling, bad actors can identify targets for nefarious behavior and infer enough demographic and personal information about them to identify them, not just on the Internet, but in real life.

As a final example, we present a hypothetical case that verges on science fiction. In November 2019, Twitter followers of the Epilepsy Foundation were targeted with posts of strobe lights and similar potentially seizure-inducing content [Aker, 2019]. The unknown actors made the assumption that people with photosensitive epilepsy would be strongly represented among this group.

Consider the scenario if those same actors had access to a highly-accurate telemedical diagnostic system that would let them identify people with epilepsy by authorship profiling. Consider that scenario, and shudder.

7 Discussion and conclusions

There is, at least in the public imagination, a pronounced tendency of scientific investigation to outstrip ethics and the public good. The vision of the Internet, as presented by organizations like the Electronic Freedom Foundation, has been replaced by one where more than half of email is spam. The primary use of digital cash is on the Dark Web. Authorship analysis technologies can address serious problems, but can also violate personal privacy at a scale never before imagined in the most dystopian of fictions.

The traditional solution to computer security problems is to attempt to educate users. Unfortunately, this simply doesn't work; as security expert Marcus Ranum has pointed out, we have been educating users for decades, and it simply hasn't worked. But the challenge posed by authorship profiling of social media is more insidious, as there's little or nothing that the user can do to prevent this kind of analysis, short of abandoning social media altogether, or of having someone else (a person or computer program) rewrite everything before it's posted. More effective action will be needed.

It may be possible to address these risks at a public policy level, but that also seems impractical. Governments' track records on dealing with cyberthreats is not good. Most practical improvements in cybersecurity have come from industry consortiums, but it's not clear that industry giants like Facebook want to acknowledge these risks, let alone fix theem, given the the need to protect their bottom lines.

The final line of defense, then, may be us – researchers in computer science, text classification, and social media analysis. These risks and problems are arguably a problem of our own creation, and we are therefore in the best place to identify and develop protections and countermeasures. The development of these protections and countermeasures should be an important research effort going forward. It is easy to build a tool to violate privacy. It is much harder, but more important, to build tools to protect and restore it.

Authorship analysis of social media is a very powerful tool with many important applications and a nearly limitless commercial potential. But we must be aware that behind the glittering gold, there may be a dragon, and we should give at least a passing thought about how to fight it.

References

[Ainsworth and Juola, 2019] Ainsworth, J. and Juola, P. (2019). Who wrote this?: Modern forensic authorship analysis as a model for valid forensic science. online at https://wustllawreview.org/essays/who-wrote-this-modern-forensic-authorship-analysis-as-a-model-for-valid-forensic-science/.

[Aker, 2019] Aker, J. (2019). Epilepsy foundation files criminal complaint and requests investigation in response to attacks on twitter feed. https://www.epilepsy.com/release/2019/12/epilepsy-foundation-files-criminal-complaint-and-requests-investigation-response.

[Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[Argamon et al., 2005] Argamon, S., Dhawle, S., Koppel, M., and Pennebaker, J. W. (2005). Lexical predictors of personality type. In *Proceedings of the Classification Society of North America Annual Meeting*.

[Argamon et al., 2009] Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

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

[Baayen et al., 2002] Baayen, R. H., van Halteren, H., Neijt, A., and Tweedie, F. (2002). An experiment in authorship attribution. In *Proceedings of JADT 2002*, pages 29–37, St. Malo. Université de Rennes.

[Bachrach et al., 2012] Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. (2012). Personality and patterns of facebook usage. In *Proceedings of the 4th annual ACM web science conference*, pages 24–32. ACM.

[Benton et al., 2017] Benton, A., Mitchell, M., and Hovy, D. (2017). Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

[Binongo, 2003] Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17.

[Bransford et al., 1972] Bransford, J. D., Barclay, J. R., and Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive psychology*, 3(2):193–209.

[Brooks, 2013] Brooks, R. (2013). Whodunnit? JK Rowling's secret life as wizard crime writer revealed. *Sunday Times*, 14 July.

[Brooks and Flyn, 2013] Brooks, R. and Flyn, C. (2013). JK Rowling: The cuckoo in crime novel nest. *Sunday Times*, 14 July.

[Cafiero and Camps, 2019] Cafiero, F. and Camps, J.-B. (2019). Why Molière most likely did write his plays. *Science Advances*, 5:eaax5489.

[Cavnar and Trenkle, 1994] Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *1994 Symposium on Document Analysis and Information Retrieval*, pages 161–176.

[Chaski, 2005] Chaski, C. E. (2005). Who's at the keyboard: Authorship attribution in digital evidence invesigations. *International Journal of Digital Evidence*, 4(1):n/a. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

[Chaski, 2007] Chaski, C. E. (2007). The keyboard dilemma and forensic authorship attribution. Advances in Digital Forensics III.

[Cohen, 2014] Cohen, N. (2014). Putative Bitcoin author categorically denies it. *New York Times*, March 17, 2014.

[Collins, 2013] Collins, P. (2013). Poe's debut, hidden in plain sight. *The New Yorker*, October.

[Corney et al., 2002] Corney, M., de Vel, O., Anderson, A., and Mohay, G. (2002). Gender-preferential text mining of e-mail discourse. In *Proceedings of* Computer Security Applications Conference, 2002, pages 282–289.

[Coulthard, 2013] Coulthard, M. (2013). On admissible linguistic evidence. *Journal of Law and Policy*, XXI(2):441–466.

[de Morgan, 1882] de Morgan, A. (1851/ 1882). Letter to Rev. Heald 18/08/1851. In Sophia Elizabeth. De Morgan (Ed.) *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters.*

[Diederich et al., 2003] Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship attribution with support vector machines. *Applied intelligence*, 19(1-2):109–123.

[Friedman and Friedman, 1957] Friedman, W. F. and Friedman, E. S. (1957). *The Shakespearean Ciphers Examined.* Cambridge University Press, Cambridge.

[Gibney, 2018] Gibney, E. (2018). The scant science behind cambridge analytica?s controversial marketing techniques. *Nature*.

[Gómez-Adorno et al., 2018] Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., and Pinto, D. (2018). Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):741–756.

[Grieve, 2005] Grieve, J. W. (2005). Quantitative authorship attribution: A history and an evaluation of techniques. Master's thesis, Simon Fraser University. URI: http://hdl.handle.net/1892/2055, accessed 5.31.2007.

[Havigerová et al., 2019] Havigerová, J. M., Haviger, J., Kučera, D., and Hoffmannová, P. (2019). Text-based detection of the risk of depression. *Frontiers in psychology*, 10.

[Herper, 2014] Herper, M. (2014). Linguist analysis says Newsweek named the wrong man as Bitcoin's creator. *Forbes Magazine*, March 10.

[Holmes and Forsyth, 1995] Holmes, D. I. and Forsyth, R. S. (1995). The Federalist revisited : New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–27.

[Hota et al., 2006] Hota, S. R., Argamon, S., Koppel, M., and Zigdon, I. (2006). Performing gender: Automatic stylistic analysis of Shakespeare's characters. In *Proceedings of* Digital Humanities 2006, pages 100–104, Paris.

[Illibagiza Umulisa, 2019] Illibagiza Umulisa, C. (2019). Author profiling, the writing style of rwandans. In *Digital Humanities and Computer Science 2019*, Chicago.

[Jockers and Witten, 2010] Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2):215–223.

[Juola, 2006] Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

[Juola, 2008] Juola, P. (2008). Authorship attribution : What mixture-of-experts says we don't yet know. In *Proceedings of American Association for Corpus Linguistics 2008*, Provo, UT USA.

[Juola, 2012] Juola, P. (2012). Large-scale experiments in authorship attribution. *English Studies*, 93(3):275–283.

[Juola, 2013a] Juola, P. (2013a). How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. *Scientific American*, August.

[Juola, 2013b] Juola, P. (2013b). Stylometry and immigration: A case study. *Journal of Law and Policy*, XXI(2):287–298.

[Juola, 2015] Juola, P. (2015). The Rowling case: A proposed standard protocol for authorship attribution. *DSH (Digital Scholarship in the Humanities).*

[Juola, 2017] Juola, P. (2017). Detecting contract cheating via stylometric methods. In *Plagiarism Across Europe and Beyond*, Brno, Czechia.

[Juola, 2018] Juola, P. (2018). Authorship attribution in a native american language (arapaho). In *Linguistic Society of American 2018 Annual Meeting*, New York City.

[Juola and Mikros, 2016a] Juola, P. and Mikros, G. K. (2016a). Authorship attribution using different languages. In *Digital Humanities 2016*, pages 241–243, Krakow.

[Juola and Mikros, 2016b] Juola, P. and Mikros, G. K. (2016b). Cross-linguistic stylometric features: A preliminary investigation. In *JADT 2016*, Nice, France.

[Juola and Mikros, 2017] Juola, P. and Mikros, G. K. (2017). Cross-linguistic correlations in lexical complexity. In *13th Biennial Conference of the International Association of Forensic Linguists*, University of Porto, Portugal.

[Juola and Noecker Jr., 2014] Juola, P. and Noecker Jr., J. I. (2014). Inferring self-esteem from keyboard behavior. In *Proceedings of DHCS 2014*.

[Juola et al., 2013] Juola, P., Noecker Jr, J. I., Stolerman, A., Ryan, M. V., Brennan, P., and Greenstadt, R. (2013). Keyboard behavior-based authentication for security. *IT Professional*, 15:8–11.

[Kernot et al., 2017] Kernot, D., Bossomaier, T., and Bradbury, R. (2017). The stylometric impacts of ageing and life events on identity. *Journal of Quantitative Linguistics*, pages 1–21.

[Khonji et al., 2015] Khonji, M., Iraqi, Y., and Jones, A. (2015). An evaluation of authorship attribution using random forests. In *2015 International Conference on Information and Communication Technology Research (ICTRC)*, pages 68–71. IEEE.

[Koppel et al., 2003] Koppel, M., Argamon, S., and Shimoni, A. (2003). Automatically categorizing written texts by author gender. *Literary And Linguistic Computing*, 17(4).

[Koppel et al., 2002] Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412. doi:10.1093/llc/17.4.401.

[Koppel et al., 2009] Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

[Kucukyilmaz et al., 2006] Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., and Can, F. (2006). Chat mining for gender prediction. *Lecture Notes in Computer Science*, 4243:274–283.

[Leonard, 2006] Leonard, R. A. (2006). Applying the scientific principles of language analysis to lssues of the law. *International Journal of the Humanities*, 3:2005.

[Luyckx and Daelemans, 2008a] Luyckx, K. and Daelemans, W. (2008a). Personae: a corpus for author and personality prediction from text. In *LREC*.

[Luyckx and Daelemans, 2008b] Luyckx, K. and Daelemans, W. (2008b). Using syntactic features to predict author personality from text. In *Proceedings of Digital Humanities*.

[Markovikj et al., 2013] Markovikj, D., Gievska, S., Kosinski, M., and Stillwell, D. J. (2013). Mining facebook data for predictive personality modeling. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[Martindale and McKenzie, 1995] Martindale, C. and McKenzie, D. (1995). On the utility of content analysis in authorship attribution: The Federalist Papers. *Computers and the Humanities*, 29:259–70.

[Matthews and Merriam, 1993] Matthews, R. A. J. and Merriam, T. V. N. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 8(4):203–209.

[McMenamin, 2011] McMenamin, G. (2011). Declaration of Gerald McMenamin. Available online at `http://www.scribd.com/doc/67951469/Expert-Report-Gerald-McMenamin`.

[Mendenhall, 1887] Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, IX:237–49.

[Merriam and Matthews, 1994] Merriam, T. V. N. and Matthews, R. A. J. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9(1):1–6.

[Mikros and Perifanos, 2013] Mikros, G. K. and Perifanos, K. (2013). Authorship attribution in greek tweets using multilevel author's n-gram profiles. In *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013, Stanford, California*, pages 17–23. AAAI Press, Palo Alto, California.

[Mosteller and Wallace, 1963] Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

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

[Noecker Jr et al., 2013] Noecker Jr, J., Ryan, M., and Juola, P. (2013). Psychological profiling through textual analysis. *LLC*, 28(3):382–387.

[Noecker Jr. and Juola, 2014] Noecker Jr., J. I. and Juola, P. (2014). Stylometric identification of manic-depressive illness. In *Proceedings of DHCS 2014*.

[O'Brien, 2018] O'Brien, S. (2018). Employers check your social media before hiring. many then find reasons not to offer you a job. *CNBC.com*, August 10, 2018.

[Osoba and Welser IV, 2017] Osoba, O. A. and Welser IV, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence. Technical Report RR-1744-RC, RAND Corporation, Santa Monica, CA.

[Pesca, 2018] Pesca, M. (2018). Stylometry for dummies: What prose analysis can (and can't) tell us about that infamous New York Times op-ed. *slate.com*, September 10, 2018.

[Rockeach et al., 1970] Rockeach, M., Homant, R., and Penner, L. (1970). A value analysis of the disputed Federalist Papers. *Journal of Personality and Social Psychology*, 16:245–50.

[Rudman, 1998] Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31:351–365.

[Rudman, 2005] Rudman, J. (2005). The non-traditional case for the authorship of the twelve disputed Federalist Papers : A monument built on sand. In *Proceedings of ACH/ALLC 2005*, Victoria, BC. Association for Computing and the Humanities.

[Scott, 2018] Scott, M. (2018). Cambridge analytica helped 'cheat' brexit vote and us election, claims whistleblower. https://www.politico.eu/article/cambridge-analytica-chris-wylie-brexit-trump-britain-data-protection-privacy-facebook/.

[Sekulić et al., 2018] Sekulić, I., Gjurković, M., and Šnajder, J. (2018). Not just depressed: Bipolar disorder prediction on reddit. *arXiv preprint arXiv:1811.04655*.

[Shuy, 1998] Shuy, R. W. (1998). *The Language of Confession, Interrogation, and Deception*. Sage, Thousand Oaks.

[Sousa-Silva et al., 2011] Sousa-Silva, R. et al. (2011). twazn me!!! : Automatic authorship analysis of micro-blogging messages. In Muoz, R., Montoyo, A., and Mtais, E., editors, *Natural Language Processing and Information Systems*, volume 6716 of *Lecture Notes in Computer Science*, pages 161–168. Springer, Berlin / Heidelberg.

[Stamatatos, 2009] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–56.

[Stamatatos, 2013] Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, XXI(2):420–440.

[Tweedie et al., 1996] Tweedie, F. J., Singh, S., and Holmes, D. I. (1996). Neural network applications in stylometry : The Federalist Papers. *Computers and the Humanities*, 30(1):1–10.

[van Halteren et al., 2005] van Halteren, H., Baayen, R. H., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

[Vescovi, 2011] Vescovi, D. M. (2011). Best practices in authorship attribution of English essays. Master's thesis, Duquesne University.

[Zhao and Zobel, 2005] Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. In *Asia Information Retrieval Symposium*, pages 174–189. Springer.