Hi Doppelgänger: Towards Detecting Manipulation in News Comments

Jan Pennekamp, Martin Henze, Oliver Hohlfeld
Communication and Distributed Systems
RWTH Aachen University, Germany
{firstname.lastname}@comsys.rwth-aachen.de

Andriy Panchenko
IT Security
Brandenburg University of Technology, Germany
{firstname.lastname}@b-tu.de

ABSTRACT

Public opinion manipulation is a serious threat to society, potentially influencing elections and the political situation even in established democracies. The prevalence of online media and the opportunity for users to express opinions in comments magnifies the problem. Governments, organizations, and companies can exploit this situation for biasing opinions. Typically, they deploy a large number of pseudonyms to create an impression of a crowd that supports specific opinions. Side channel information (such as IP addresses or identities of browsers) often allows a reliable detection of pseudonyms managed by a single person. However, while spoofing and anonymizing data that links these accounts is simple, a linking without is very challenging.

In this paper, we evaluate whether stylometric features allow a detection of such doppelgängers within comment sections on news articles. To this end, we adapt a state-of-the-art doppelgänger detector to work on small texts (such as comments) and apply it on three popular news sites in two languages. Our results reveal that detecting potential doppelgängers based on linguistics is a promising approach even when no reliable side channel information is available. Preliminary results following an application in the wild shows indications for doppelgängers in real world data sets.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; • **Security and privacy**; • **Computing methodologies** → *Machine learning*;

KEYWORDS

online manipulation; doppelgänger detection; stylometry

ACM Reference Format:
Jan Pennekamp, Martin Henze, Oliver Hohlfeld and Andriy Panchenko. 2019. Hi Doppelgänger: Towards Detecting Manipulation in News Comments. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316496

1 INTRODUCTION

In the recent past, a new trend of opinion manipulation emerged on the Internet. As a consequence, media outlets increasingly report about fake news, propaganda, and trolls (online identities manipulating other users with their postings) [23]. While most of these concepts are well-known from a pre-Internet era, they have become more dangerous than before as a significant amount of users relies on the Internet as their primary information source [14]. Consequently, the validity of online information is at stake, especially since users are ubiquitously confronted with user-generated content, e.g., in forums, social networks, and comments to news articles. Governments, organizations, and companies might exploit this situation to manipulate users' opinions in their favor [34].

Traditionally, opinion manipulation on the Internet has spread through social media [4], e.g., during the Ukraine crisis [25] or the 2016 US presidential election [7]. For example, Facebook reported on closing multiple 10 000 accounts in a single week due to suspicious activities [15]. Likewise, Twitter identified more than 50 000 manipulating accounts during the presidential election [30]. Recently, we observed a shift of opinion manipulation towards the comment sections of news sites [21]. This is an especially alarming trend, since over 90 % of US Americans retrieve news online and 63 % of online news papers allow comments on articles [10]. For example, *the Guardian* reports that their moderators identify up to 250 opinion manipulating comments under a *single* article [12].

To counter the immediate threat to society introduced by online manipulation, related work usually relies on manual flagging by ordinary users (e.g., [20, 28]) or side channel information, such as IP addresses [19]. Clearly, the latter can be circumvented by applying anonymization techniques. In contrast, we focus on a scenario where manipulators actively *try* to conceal their identity, e.g., by hiding IP addresses or using fake email accounts.

In such a setting, one important building block to detect misinformation in the comment sections of news sites is the ability to detect *doppelgängers*, i.e., people using multiple identities with the goal to amplify the influence of the opinion they spread. As these trolls try to remain undiscovered, side channel information, such as IP addresses or email accounts, is often not reliable [17]. Thus, doppelgänger detection has to mainly rely on information which is difficult to spoof. While anonymizing or spoofing technical communication aspects (e.g., IPs) is feasible, consistently maintaining separate writing styles is hard [2]. In this regard, we explore the applicability of characteristics of written texts to facilitate this task. Detecting doppelgängers based on written text has been studied before for longer texts, such as articles or blog posts [24]. However, comments on news sides are considerably shorter, rendering detection of doppelgängers much more challenging.

In this paper, we target the question on whether detecting doppelgängers in comment sections of news sites is feasible merely by leveraging linguistic features. Ultimately, this approach assists

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316496

providers of news sites to detect attacks on the public opinion even if manipulators purposely disguise their behavior by obfuscating client-specific data, such as IP addresses, used browsers, or email addresses. More precisely, our contributions are as follows:

(1) We extend a state-of-the art doppelgänger detector [3] with additional features to account for comment-specific characteristics (short and informal text with only 50 to 1 000 characters).
(2) We evaluate the feasibility of doppelgänger detection for short news comments solely based on linguistic features using a random subset of more than 4.8 million authentic news comments from three popular news sites in two languages. Our results show that our approach is able to detect the presence of doppelgängers with high detection accuracy, even when the number of doppelgängers is unknown. It also works well on two languages, German and English.
(3) We briefly hint at preliminary results of applying our approach to detect doppelgängers in the wild. Here, we find indications for user accounts constituting doppelgängers.

With our work, we contribute a building block to methods for fighting public opinion manipulation in online media.

2 RELATED WORK & BACKGROUND

To remedy the immediate threat resulting from opinion manipulation, especially in the context of online news comments, providing approaches that reliably *detect* opinion manipulation is imperative. Only if we are able to detect opinion manipulation, we can take appropriate countermeasures, such as flagging, moderating, or deleting comments. In the following, we first provide an overview of different approaches to detect opinion manipulation on the Internet. Subsequently, we take a closer look at approaches to detect doppelgängers, i.e., opinion manipulators who utilize more than one user account to amplify the credibility of their agenda, and especially those approaches that focus on stylometric features.

2.1 Detection of Opinion Manipulation

The first approaches tackling the detection of opinion manipulation were concerned with identifying spam in product reviews (e.g., [6, 19]). Subsequent research shifted towards social networks. E.g., Vosoughi et al. [31] conduct a longitudinal analysis of news story spreading on Twitter while Zannettou et al. [33] additionally include Reddit and 4chan into their analysis of manipulation.

On a related note, Zannettou et al. [32, 34, 35] investigate different aspects of the behavior of government-paid trolls. Kumar et al. [18] analyze the behavior of trolls in forums based on IP addresses and which discussions specific accounts participate in. Mihaylov et al. [20] expose paid trolls in comment sections based on their commenting behavior. Subsequently, they extend their approach [22] to also incorporate linguistic features to discover paid trolls. However, they only target spamming or annoying trolls. Consequentially, their approach would likely fail to discover user accounts with the focus of spreading misinformation.

2.2 Detection of Doppelgängers

To detect opinion manipulation, especially in the context of news sites, one important building block is the ability to detect doppelgängers, i.e., manipulators who operate multiple accounts to increase the impact of the opinion they express in their comments. Existing approaches to detect doppelgängers can be coarsely classified into approaches that either rely on metadata (e.g., time stamps or the topics that accounts are active in) or stylometric features (e.g., lexical, syntactic, and domain-specific writing style characteristics).

Using Metadata to Detect Doppelgängers. The most intuitive approach to detect doppelgängers relies on metadata such as time stamps of the published messages. Zheng et al. [36] detect doppelgängers in forums based on their activity frame and the topics they posted messages in. Johansson et al. [17] introduce the notion of "timeprints" to detect doppelgängers in social networks. Such a timeprint describes the account's activity pattern in a fixed set of features. The authors claim that the activity patterns are similar across accounts handled by the same manipulator.

Using Stylometry to Detect Doppelgängers. In contrast to approaches that rely on metadata, stylometric approaches focus solely on the written text and strive to identify text snippets that originate from the same author. This difference becomes especially relevant as manipulators try to avoid detection [17], hence rendering metadata, e.g., activity patterns, useless.

As a foundation to detect doppelgängers solely based on their written text, all approaches rely on so-called *stylometric features*, i.e., information that can (easily) be extracted from written text that, ideally, is unique for each individual author. Abbasi et al. [1] defined the most-relevant set of stylometric features, called *writeprints*, to model all aspects of writing style, especially targeting programming code snippets, ebay buyer/seller feedback, emails, and chat messages. These stylometric features cover lexical, syntactic, structural, content, and idiosyncratic aspects of written text. To also account for special characteristics of modern online communication, related work proposed to extend the writeprints set with additional domain-specific features, such as the sentiment of words [11], the number of received votes [20], and the usage of emoticons [11]. Authorship attribution is a well-established field of research and related work achieves promising results on longer texts, such as book chapters and blog posts [29]. Only recently, focus shifted towards shorter messages (as they are prevalent in news comments), e.g., in the context of Twitter [8, 26]. However, when considering news comments, existing approaches only rely on stylometric aspects to a very limited extent [22].

Using different sets of features, approaches from related work apply stylometry to detect doppelgängers in different settings and types of text. For example, Almishari et al. [5] demonstrate the linkability of Twitter accounts over time based on two simple features. They only consider unigrams and bigrams (n-grams measure the frequency of character sequences) to match doppelgängers. Afroz et al. [3] compile a subset of lexical and syntactic features from the writeprints feature set and extend it with a feature expressing the leetspeak percentage to detect doppelgängers in blogs and forums. They input their features into the Doppelgänger Finder algorithm.

Doppelgänger Finder. This algorithm [3] is an approach to compare the writing styles of different authors with each other to obtain a similarity score for each pair of authors. The algorithm consists of three steps: (1.) selecting meaningful features from the complete set of features by conducting a principal component analysis (PCA), (2.) applying logistic regression to calculate the similarity of the writing style to all other authors for each author, and (3.) combining

these similarity scores to a single score for each pair of authors. To interpret these scores, a threshold depending on various properties of the dataset [3] must be applied. Afroz et al. [3] recommend a manual analysis to identify such a threshold.

The approach taken by Doppelgänger Finder is especially relevant for our work, as the general algorithm is applicable to different means of communication. Furthermore, its implementation is freely available and thus an ideal foundation for our work. More specifically, in this work, we extend the overall process of comparing stylometric features using Doppelgänger Finder with additional features and an adaptable threshold metric to apply it for detecting doppelgängers in news comments, which often are shorter than postings in blogs and forums.

3 DETECTING DOPPELGÄNGERS IN NEWS COMMENTS

To detect doppelgängers in news comment sections, we present an approach which takes a set of user accounts, each having a pseudonym or user name, as input and generates a list of potential pairs of doppelgängers within this set of accounts as output. As such, we provide a check whether a single individual publishes texts with multiple accounts solely based on stylometric information. Our approach consists of three building blocks: (1.) an augmentation of a state-of-the-art stylometric feature set [1] to adapt to modern online communication, (2.) an application of a state-of-the-art doppelgänger detector [3] to retrieve similarity scores for each pair of accounts, and (3.) an automated threshold metric to categorize the obtained similarity scores into potential doppelgängers and regular accounts. We now present each of these building blocks in detail.

3.1 Stylometric Features for News Comments

As a foundation to detect doppelgängers based on written text, we first craft a stylometric feature set. We augment features proposed by different approaches from related work [11, 20] with own custom features specific to comments in news sections and create a unique feature set that – as the evaluation shows – is efficient in our targeted scenario.

Writeprint. We rely on the unmodified extended writeprints feature set consisting of 21 categories, such as word length distribution, part-of-speech tags, and word n-grams. Their default set of more than 27.6 k features (as we could not reproduce the 2 300 part-of-speech tags, we follow the approach of Doppelgänger Finder [3] with 45 unigrams) is enlarged by the combination of occurring bigrams and trigrams. By default, we include all combination which accounts for more than 93 k additional features. While Doppelgänger Finder [3] uses only a reduced set of 8 categories ($\approx$ 1.1 k features), we implement the complete set.

Adaption for the Web. To adjust this feature set to the characteristics of modern online communication, we also include various extensions from related work [11, 20]. In particular, we decided to include an additional set of 165 features (f.) expressing: (i) the sentiment of comments [11] (4 f.: positivity and sensitivity per word and sentence), (ii) the activity periods of users [20] (4 f.: weekdays or weekends, working hours or night hours), (iii) the number of received votes [20] (7 f.: histogram with 7 intervals), and (iv) the usage of emoticons [11] (150 f.: custom list of 150 emoticons).

Our Additions. Furthermore, we calculate additional idiosyncratic features (8 f.: frequency of grammar mistakes and uppercase word usage per sentence and comment) to also cover more user-specific writing style habits. We additionally measure the amount of whitespace (2 f.: per sentence and message) and the comment length (2 f.) as well as the use of 100 popular hashtags (100 f.), reply functionality (1 f.), quotations (3 f.: usage and placement), and 50 popular link shorteners (50 f.) to cover aspects specific to news comments.

In total, our stylometric feature set consists of 121 378 features. While most of these features can be applied for every news site, some specific features, e.g., quotations from previous comments, can only be used if a site implements the functionality. Recall that Doppelgänger Finder applies a PCA for dimension reduction to the feature set, making the effectively used features dependent on the training data. The analysis of feature relevance is beyond the scope of our feasibility study, which is why we leave it to future work.

All these features operate solely on public information, i.e., no private side channel information, such as IP addresses or email addresses, is used. The rational here is that modifications to the comment with the goal to avoid detection of manipulation are unlikely as any changes to the wording can have an impact on the success of the attempted opinion manipulation.

3.2 Applying Doppelgänger Finder to News Comments

To apply Doppelgänger Finder (cf. Section 2.2) to news comments, we first need a set of user accounts (pseudonyms) for our analysis. Then, we need to extract the writing style features (using our stylometric feature set) from a user's comments (with a certain minimum text length). To extract our features, we make use of several comments written under the same account, following the stylometric evaluations of Twitter [8]. We then input these stylometric profiles into Doppelgänger Finder to identify authors with similar writing styles (potential doppelgängers). More specifically, Doppelgänger Finder computes a similarity score for each pair of authors.

To categorize an author pair as either doppelgänger or non-doppelgänger pair, we have to define a threshold to provide a semantic to our evaluation results, i.e., signal that the respective two author profiles are likely to belong to a single user. Doppelgänger Finder [3] does not provide a structured approach to obtain a threshold. Instead, the authors recommend a manual analysis to determine the threshold, which is tedious and potentially error-prone in scenarios with ambiguous writing styles. In this work, we present a metric to *automatically* derive such a threshold.

3.3 Automated Threshold Metric

As manual analysis to derive the best threshold is very time consuming and prone to errors, we propose an automated metric. The current set of parameters, such as the number of user accounts, the considered text length per account, and the dataset, defines our *target* scenario. Given that the threshold depends on these parameters, we *simulate* scenarios which are similar to our current *target* scenario to obtain a threshold which is generally applicable to scenarios with identical parameters. Here, the overall concept is to rely on artificially splitting user accounts, i.e., constructing known pairs of doppelgänger, to create a new controlled scenario with the same

Dataset	# Articles	# Users	# Comments	Comments / Article*[⌀]	Users / Article*[⌀]	Comments / User [⌀]	Length / Comment [⌀]
The Guardian	210 819	48 033	3 023 674	385.1	85.9	62.9	262.7
SPIEGEL ONLINE	89 480	11 460	808 245	62.2	35.1	70.5	505.8
ZEIT ONLINE	29 784	10 265	1 057 175	171.1	67.6	103.0	434.6

* If the article was published in July or August 2017 and the comment section of the article was enabled.

Table 1: The posting behavior of user comments on news depends on the crawled website. As expected (cf. linguostatistics), German language sites on average comprise of longer comments than comments on the English language site *the Guardian*.

parameters. When splitting a user account (or pseudonym) P, we divide all comments of P randomly into two sets (P_1 and P_2), while making sure that all comments on an article belong to the same set. Instead of evaluating a user account P, we analyze comments written by its respective partitioned pseudonyms P_1 and P_2.

With this approach, we establish ground truth on our *simulated* scenarios (as we know that P_1 and P_2 are doppelgängers), and thus can group all pairs into two classes (doppelgänger and non-doppelgänger pairs). By applying a threshold, we can compare the classifier's output (i.e., similarity scores) to the known grouping of all pairs to assess the classification accuracy. Hence, we can choose the threshold such that we achieve the optimum in terms of precision and recall. Repeating these *simulated* scenarios multiple times (for our *target* scenario) allows us to automatically determine a generally applicable threshold for these particular parameters.

Finally, we can apply this threshold to our *target* scenario (without ground truth). Author pairs with a similarity score (as computed by Doppelgänger Finder) exceeding the threshold are flagged as doppelgängers, while the remaining pairs are considered to be non-doppelgängers. As this approach only provides a *heuristic* for doppelgängers, we recommend to verify this flagging through other means (either automated monitoring or manual analysis).

By crafting a stylometric feature set for news comments and extending Doppelgänger Finder with a metric for automated threshold selection, we provide an approach that allows to check arbitrary sets of author profiles on news sites for doppelgängers, i.e., single individuals publishing comments under different accounts, solely based on information directly derived from the published comments.

4 DATA SET

We evaluate our approach by applying it to detect doppelgängers in user comments on online news sites. Since no public data set of texts published in comment sections exists, we obtain 5 million comments posted in response to 300 k articles in 3 major news papers by crawling. We next describe our crawling approach and the resulting data set.

4.1 Selection Criteria

For our evaluation, we want to target popular news sites that can be of high relevance to opinion manipulation by the ability to reach a large target audience. Hence, we focus on popular news sites (according to their Alexa rank) with an ability for users to post comments. For crawling, we require sites to display comments on their own site, i.e., without embedding third-party components such as *Facebook* or *Disqus*. To study doppelgänger accounts, we require sites to offer a proper authentication mechanism to ensure that all comments are uniquely assigned to a user profile or account number in the data set. Otherwise, displayed author names do not necessarily refer to the same individual and thus might already represent a mixture of writing styles from multiple authors, even without manipulative doppelgängers. Furthermore, dedicated profiles allow us to list all comments published by a single user, easing the collection of a large number of comments for each profile. In our opinion, this constraint does not limit the applicability of our approach because active, connected, and long lasting user accounts are more likely to convince users when spreading misinformation.

Our final selection consists of *the Guardian* (2nd most popular news website in UK), *SPIEGEL ONLINE* (most popular news website in DE), and *ZEIT ONLINE* (5th most popular news website in DE). While selecting pages with different languages helps us to check the language independence of our approach, choosing two pages in the same language allows us to investigate differences between news platforms in the same country.

Only relying on public comment data that is available through crawling forms a minimal set of information that exacerbates the detection problem. That is, platform providers can still enrich our approach by using further side channel information or non-public user account identifiers to map multiple comments to a single pseudonym. Such a proprietary data set is, however, only available to the platform providers themselves, while our approach can be directly applied by the general public.

4.2 Crawling Period

We periodically crawled the selected news sites over a period of two month (July 1, 2017 to August 31, 2017) to collect large comments data sets. This period covers two rounds of United Kingdom's European Union exiting negotiations, which might be a potential target for opinion manipulation and spreading of misinformation.

During our crawling period, we scanned the main page of each news site every 30 s for new articles. For each article, we retrieved all comments at different intervals (after 1 h, 4 h, 12 h, 24 h, 48 h, and 7 d). The motivation here is that our data set will also contain altered comments (moderated, edited, or deleted). Besides, our assumption is that comment sections are most active after an article has initially been published and consequentially, early published comments might cause the most severe damage wrt. opinion manipulation. Finally, after the two month period, we collected the 100 most recent comments per user account (for *the Guardian*, we only processed 62 % of the users as they asked us to cease our crawling activity).

4.3 Data Set Statistics

We retrieved ≈ 5 million authentic comments of nearly 60 000 user accounts on over 300 000 news articles as summarized in Table 1.

Our data sets also record website-specific features, such as the use of quotations (*the Guardian* and *SPIEGEL ONLINE*), the number of votes (*the Guardian* and *ZEIT ONLINE*), or comment headlines (*SPIEGEL ONLINE*). Besides, we also observe that moderation takes place on the crawled sites (e.g., user accounts being closed or comments being modified/deleted). Although we refrained from assessing the reasons for these actions, misconducts, such as defamation, harassment, spam, or opinion manipulation, are likely reasons.

5 EVALUATION

We evaluate our approach by artificially splitting user accounts (cf. Section 3.3) and thereby *simulating* the presence of doppelgängers in real-world texts. This method is necessary since we have no ground truth data set with labeled doppelgängers. By artificially splitting accounts, we assess the ability of our approach to detect the splitted accounts by the same author, and thereby detecting doppelgängers.

5.1 Simulation of Doppelgängers Approach

Our evaluation approach follows established prior work (e.g., Abbasi et al. [1] or Johansson et al. [16]) to split user accounts into pseudonyms. This way, we introduce *known* doppelgängers into real-world texts without requiring labeled data of doppelgängers. We deliberately decided to experiment on real-world texts instead artificially created texts (e.g., written by paid *Amazon Mechanical Turk* workers [13]) to test our approach on realistic input data.

In our evaluation, we conduct a 3-fold cross-validation for each experiment. First, we randomly select user accounts from our data sets and artificially split them as described previously (acting as pseudonyms). Then, we apply our approach (cf. Section 3) to each fold. Finally, we average the results over all three folds. Note that our results correspond to a lower bound because the random set of user accounts could already contain doppelgänger pairs which we are unaware of due to the missing ground truth.

5.2 Applied Statistical Measures

The traditional definition of precision and recall is not applicable to all evaluation scenarios because the number of included artificial doppelgänger pairs can also be zero (varying number of doppelgängers). Hence, given that we have no true positives, we would always end with a precision and recall of 0 %. An alternative definition of counting the correctly predicted pairs is unsuitable as well because in larger scenarios the non-doppelgänger pairs (increasing quadratically) dominate the result (the number of evaluated accounts only scales linearly), i.e., reducing the use of precision and recall to absurdity. Consequentially, we need a custom definition that is unaffected by the number of included doppelgängers while not being dominated by the majority of non-doppelgänger pairs.

Our following custom definition of the entries in a confusion matrix addresses these issues. Hence, we can compare the results of different evaluations regardless of the adjusted parameter. However, a drawback of this adjustment is that in scenarios with only a few doppelgänger pairs, missed pairs do not result in a low accuracy. In our opinion, this limitation is not an issue because we want to focus on reliably detecting these doppelgänger pairs without introducing incorrectly flagged pairs. Nevertheless, as part of our evaluation, we conduct experiments with varying numbers of doppelgänger pairs to cover different scenarios.

For each pseudonym, we refer to the pairs that the classifier correctly predicts to be a doppelgänger as $\mathrm{cDG_p}$ (correct Doppelgänger pairs). We count the pairs that the classifier incorrectly predicts to be a doppelgänger as $\mathrm{FP_p}$ (False Positive pairs) and the pairs that the classifier fails to predict to be a doppelgänger as $\mathrm{FN_p}$ (False Negative pairs). The lower $_\mathrm{p}$ denotes that we count the number of pairs (each pseudonym is compared to all others). Next, we introduce our mapping from these pair counters, i.e., $\mathrm{cDG_p}$, $\mathrm{FP_p}$, and $\mathrm{FN_p}$, to entries in a confusion matrix for each pseudonym.

$$\mathrm{TP} = \begin{cases} \dfrac{\mathrm{cDG_p}}{\mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p}} & , \text{ if } \mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p} \neq 0, \\ 1 & , \text{ otherwise.} \end{cases}$$

$$\mathrm{FP/FN} = \begin{cases} \dfrac{\mathrm{FP_p}/\mathrm{FN_p}}{\mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p}} & , \text{ if } \mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p} \neq 0, \\ 0 & , \text{ otherwise.} \end{cases}$$

We apply the commonly known definitions of precision and recall to these mappings and present the results as weighted F-score ($F_{0.5}$) that favors precision over recall [27] as we are more interested in correctly detecting pairs of doppelgängers than detecting all doppelgängers (while accepting incorrectly flagged pairs).

5.3 Known Number of Doppelgängers

We begin with evaluating a *baseline scenario* in which we use a-priori knowledge on the number of doppelgängers to correctly set the threshold. We omit this baseline setting in the next section where we do not use a-priori knowledge for a more realistic setting.

Approach. We base our decisions and initial experiments on our *the Guardian* dataset. Our approach requires three different parameters to be set in advance: (i) the number of accounts that we want to evaluate, (ii) the number of instances that represent the writing style, and (iii) similarly, the text length that a single instance consists of. Overall, we select realistic values that conform to our *the Guardian* dataset. For the number of accounts, we choose a total of 100 pseudonyms since this number exceeds the average number of 85 user accounts commenting on an article (cf. Table 1). Second, this setting still covers 77.67 % articles in our *the Guardian* dataset.

Given the shorter length of news comments compared to forum postings originally used to develop doppelgängers finder [3], we double their choice and select 20 instances per pseudonym. This number of writing style profiles, which we generate per pseudonym, correlates with information that we have available for each pseudonym. The text length per instance determines the quality of a single writing style profile.

Since comments are rather short, we append multiple comments to a longer comment [8] and evaluate two different strategies to identify the optimal decision; we can either take a fixed number of comments, or we can require a minimum number of characters.

Results. We show the doppelgängers detection accuracy as $\mathrm{F_{0.5}}$ in Table 2 (row: Complete Scenario) for randomly selected comments with a minimum appended length of 250 to 1 000 characters and 2 to 6 randomly selected (and appended) comments. We selected these increments because the average length of a comment on *the Guardian* is 262.70 characters. We refrained from evaluating a single comment because this randomly chosen comment can

Scenario	Length [char.]				Comments [#]		
	250	500	750	1 000	2	4	6
● *Complete*	95.11	98.74	99.32	99.73	94.53	98.61	97.13
◐ *Partial*	95.92	97.40	99.46	99.57	93.22	97.57	95.30

Table 2: *Simulated Doppelgängers:* Results for *the Guardian* dataset scenarios with 20 instances per pseudonym and a total of 100 pseudonyms depending on the composition of an instance. We split 50 user accounts into two pseudonyms each and list the $F_{0.5}$-scores measured in percent [%].

be significantly shorter than the average. The results highlight a high detection accuracy in every scenario. Further, more textual information improves the classification accuracy.

To increase the applicability (trade-off between accuracy and required comments per pseudonym), we fix the text length per instance at >750 characters because the improvements saturate, i.e., our approach correctly separates doppelgängers and non-doppelgängers for most of evaluated pseudonyms. With these parameter decisions, we require approximately 60 comments for every pseudonym that is part of our evaluation. In our *the Guardian* dataset, this corresponds to more than 99.48 % of all user accounts. Furthermore, we believe that pseudonyms with more published comments have more influence (wrt. opinion manipulation) because they appear more trustworthy than fresh accounts. Hence, we expect that we are able to collect a sufficient number of comments for every pseudonym that is relevant for doppelgängers checks.

When comparing this decision to related work, Afroz et al. [3] train their classifier on at least 4 500 words per author. When applying an average word length, such as 5.10 characters [9], we end up with approximately 22 950 characters per author. With our setting of 750 characters per instance and 20 instances per author, we end up with a lower bound of only 15 000 characters. Consequentially, our *baseline scenario* maps to our setting of shorter targeted texts and provides initial reasonable results to serve as a candidate selection for doppelgänger detection.

Comment Variability. The previous evaluations relied on selecting a fixed, randomly selected set of comments per author and thus did not enable us to assess the influence of variability between different comments on the detection accuracy. To assess the influence of this variability, we now repeat the previous evaluations but multiple times (over 5 runs). In each run, we select different comments from each author. Overall, the variance in the results following randomly chosen comments is marginal and ranges between 4 to 8 incorrectly classified pseudonyms out of 300 (3 folds with 100 pseudonyms each). We thus conclude that the influence of comment variability is insignificant.

Restricted News Resorts. Next, we check whether the content of our evaluated comments has an impact on the classification accuracy. We therefore select comments from articles in critical resorts (politics, news, education, and environment) on *the Guardian* that might be more prone to the spreading of misinformation. We list the respective results in Table 2 (row: Partial Scenario). We are unable to observe a definite trend in either direction and therefore, we assume that our approach is feasible for the entire platform of *the Guardian*, irrespective of the comments' content. Consequentially, we conclude that the content of comments has a negligible influence on our proposed approach.

Having analyzed the influence of the text length per instance, we next focus on the remaining two parameters: number of instances and number of pseudonyms. We stick to our *baseline scenario* and only adjust a single parameter at once.

Number of Instances. An increasing number of instances increases the classification accuracy ($F_{0.5}$-score of 99.59 % for 25 instances and 100.00 % for 30 instances). Contrary, too few instances per pseudonym significantly impair the results (e.g., $F_{0.5}$-score of 90.98 % for 10 instances and only 77.43 % for 5 instances). This decision must be made in light of the trade-off between accuracy and required comments per pseudonym. If we analyze more comments per user account, we can create additional stylometric profiles, i.e., differentiate them better from other unrelated profiles. However, we might not have a sufficient number of comments for every account.

Number of Pseudonyms. Adjusting the number of pseudonyms per account provides similar results in scenarios with 150 and 200 pseudonyms ($F_{0.5}$-score of 98.81 % and 98.25 %, respectively). Increasing the parameter even further to 250 and 300 pseudonyms, results in a decline to 97.03 % and 96.76 %. This observation is reasonable because the number of pairs that the doppelgänger detector considers increases quadratically (as the similarity score is calculated for each pair). With an increasing number of considered writing styles, differentiating and matching the writing styles is more challenging. Nevertheless, for reasonable numbers of pseudonyms, our approach identifies most doppelgänger pairs without labeling non-doppelgängers as doppelgängers. Based on our dataset knowledge (an average of 85 user accounts commenting on an article), we can conclude that our proposed approach fits the targeted scenario.

Findings. *The baseline scenario with a-priori knowledge on the number of doppelgängers provides high detection results for realistic comment lengths, the number of comments, and pseudonyms per author.*

5.4 Unknown Number of Doppelgängers

We further evaluate a more realistic scenario in which the number of doppelgängers is *unknown* and thus the threshold cannot be set based on the number of doppelgängers. In the following, we show that we achieve comparable results regardless of the number of doppelgängers. We label these experiments as *oblivious* because we do not benefit from knowledge on the expected number of doppelgängers in our simulated scenarios. This adjustment is essential as we are unaware of the number of doppelgängers in a real-world evaluation. We define three new settings: (i) *None*: our evaluation contains no simulated doppelgänger, (ii) *Single*: our evaluation contains only a single doppelgänger pair, and (iii) *Random*: we have between 25 % and 75 % doppelgängers. For comparison, we label our baseline experiment as *matching* since here the threshold is derived in simulated scenarios where the number of doppelgängers matches the experiment (known number of doppelgängers).

For our *matching* experiments, we determine the thresholds on related scenarios (as conducted before when splitting each user account into two pseudonyms). Consequentially, for the *None* setting, we are unable to determine a threshold without any simulated doppelgänger. A threshold of 1 would always yield perfect results. In *oblivious* experiments, we determine the threshold on a large

Dataset	Doppelgänger	None	Single	Random
The Guardian	*matching*	n/a	99.46	98.91
	oblivious	99.46	99.46	99.31
SPIEGEL ONLINE	*matching*	n/a	99.32	98.20
	oblivious	98.39	97.42	98.04
ZEIT ONLINE	*matching*	n/a	99.33	96.72
	oblivious	97.85	98.79	96.91

Table 3: Comparison of scenarios depending on the number of simulated doppelgängers across all datasets. We list the $F_{0.5}$-scores measured in percent [%].

number of scenarios that equally consist of all our defined settings. Hence, we do not favor a specific quota of doppelgängers when determining the threshold. As stated before, additional scenarios improve the accuracy of the threshold as it depends on the dataset.

In Table 3, we list the respective evaluation results. We again observe a high detection accuracy for all scenarios even if the number of doppelgängers is unknown to the algorithm (oblivious). Compared to the baseline setting in which the number of doppelgängers is known (matching), the detection accuracy is only slightly lower. The slight variation in the *Random* setting results from a varying number of simulated doppelgängers across multiple runs.
Findings. *Our approach is suitable for real-world settings in which the number of doppelgängers is unknown as it still provides a high detection accuracy.*

5.5 Language Dependence

To study the influence of language on the detection accuracy, we now apply our approach to the German data sets (i.e., *SPIEGEL ONLINE* and *ZEIT ONLINE*) by repeating the experiments we presented for our *the Guardian* dataset in Section 5.4.

As listed in Table 3, the results on comments written in German are slightly worse when compared to our *the Guardian* evaluation (regardless of *matching* or *oblivious* experiments). However, the general performance is similar to *the Guardian*. We expect that this slight decrease in accuracy between English and German texts probably results from the longer average text length in German (due to longer words). Repeating the steps that we conducted for *the Guardian* in Section 5.3 to determine the "text length per instance" parameter should improve the presented results marginally. Nevertheless, the results show that our general approach is applicable to German comments without any adjustments.

When comparing the German data sets to each other, we notice minimally worse results on *ZEIT ONLINE*. We believe that this trend results from two aspects. First, the conversations in *ZEIT ONLINE*'s comment sections are more colloquial when compared to *SPIEGEL ONLINE*. Second, as listed in Table 1, the average comment length on *ZEIT ONLINE* is shorter. Hence, the extracted writing style profiles are less accurate. In this case, adjusting the "text length per instance" parameter might improve the achieved results as well.

Adding further languages requires tool support (part-of-speech tagger) and structural information (e.g., word separators) on the target language to properly implement the features.
Findings. *Our approach is able to reliably detect doppelgängers in both languages (English vs. German).*

6 REAL-WORLD IMPRESSIONS

We now apply our approach in a real-world setting (without ground truth). More specifically, we conduct tests similar to what could actually be performed by news sites to detect online manipulation; we manually inspect the doppelgängers detected by applying our approach. Even in the absence of ground truth, the motivation of our anecdotal analysis is to get a first intuition on whether the detected pairs of doppelgängers are realistic.

To limit the manual effort required for checking detected doppelgängers, we restrict our analysis to 10 comment sections per website and only compare user accounts that comment on the same article. For each article, we compare a random subset of 100 user accounts to each other. We use the same parameters as in our baseline scenario (cf. Section 5.3): 20 instances per user and at least 750 characters per comment. For each data set, we use the respective threshold (t) as derived in Section 5.4.

The Guardian. In the comment section of an article demanding politicians to investigate the funding of (Islamic) extremists in Great Britain, our approach detects two accounts with a similarity score of 0.0218 (t = 0.0140). The two accounts were registered within half a year (2012/2013) and published more than 3 200 comments. Manual inspection indicates that writing styles and behaviors, e.g., preferred articles, overlap. Furthermore, both accounts oppose Brexit.

Commenting users of an article where a man is reportedly urged to take a citizenship test despite born and living in Great Britain for his whole life lead to a pair of potential doppelgängers with a similarity score of 0.0161 (t = 0.0140). Both accounts have about 210 comments but were registered in different years (2010 vs. 2016). Manual investigation reveals that both authors write long comments. However, while one frequently inserts line breaks, the other account publishes long paragraphs. Hence, without additional information, assessing whether these two accounts are indeed doppelgängers is difficult.

SPIEGEL ONLINE. Analyzing user accounts commenting on an article about an anti-terror demonstration in Cologne, Germany, our approach identifies a doppelgänger pair with a similarity score of 0.0220 (t = 0.0160). While one of the accounts has more than 1 700 comments and was registered in 2008, the other one is only rarely used for commenting (around 80 comments) since its registration in 2015. Both accounts mainly comment on articles about politics; the account with more comments is also active in resorts that are less frequently accessed, e.g., culture, panorama, and health. An additional manual analysis confirms that the two writing styles are very similar. Furthermore, both accounts frequently use comment titles, an optional feature only offered by *SPIEGEL ONLINE*.

ZEIT ONLINE. In the comments to an article on the European Union's Brexit demands, our approach uncovers a (potential) doppelgänger pair with a similarity score of 0.4445 (t = 0.0120). Manual analysis shows that both accounts publish English-language comments on a German-language news site. As English writing styles strongly differ from German writing styles, such a result is not surprising. Noticing clear differences in the command of the English language, we strongly believe that these authors are not real doppelgängers.

The same article reveals an interesting situation with two additional, overlapping pairs of (potential) doppelgängers. Authors A

and B have a similarity score of 0.0203 (t = 0.0120), while A and C have a similarity score of 0.0177. The similarity score of the third pair, i.e., B and C, is slightly below the threshold with 0.0111. To further investigate these potential doppelgängers, we cannot access the registration dates (not publicly provided by *ZEIT ONLINE*), and hence have to find other indicators to infer their actual relationship. Considering their activity periods, we find that A is active throughout the day, including night hours. A similar observation holds for B, while C is only active during typical office hours. Likewise, we observe that the writing styles of A and B are more alike than their writing styles compared to C. As supported by the similarity scores, we thus believe that A and B indeed might be doppelgängers, whereas A and C as well as B and C are likely not doppelgängers.

Overall, our anecdotal analysis detected some interesting pairs of *potential* doppelgängers. Our manual investigation is, however, unable to assess the pairs with certainty given the lack of ground truth. We contacted the respective German news sites to inquire about these potential doppelgängers, but did not receive replies. Interestingly, all but one of the detected user accounts published comments which are positioned at a prominent location directly below the text of corresponding news articles. These prominent comments are additionally indexed by the Google search engine, further increasing their potential impact on public opinion. A relevant direction for future work is therefore to apply our approach to a dataset with known doppelgängers to evaluate its performance against ground truth data. Since such a data set is unavailable to us, we leave this analysis for future work.

7 CONCLUSION

In this paper, we presented an automated approach to detect doppelgängers in comment sections of news websites based on stylometric information. In contrast to approaches that rely on technical metadata (such as IP addresses or browser identifiers)—which can be easily spoofed—our doppelgänger detection makes use of comparison of authors' "fingerprints" derived from their writing style. By using artificially splitted accounts from real-world data sets, we showed that our method is efficient in linking them together for different practical application scenarios. Our preliminary results of an application in the wild showed indications for doppelgängers in user comments of popular online news websites. These findings need to be further investigated and manually inspected to draw final conclusions about the limits of a practical applicability of our method. With our work we want to encourage the community of researchers to further investigate this topic as public opinion manipulation is a serous threat to our society with a potential to cause a severe negative influence on different spheres of our life, even in established democracies.

ACKNOWLEDGMENT

The authors would like to thank Afroz et al. [3] for making their *Doppelgänger Finder* implementation publicly available. This work has been funded by the Excellence Initiative of the German federal and state governments.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM TOIS* 26, 2.
[2] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *IEEE S&P*.
[3] Sadia Afroz, Aylin Caliskan Islam et al. 2014. Doppelgänger Finder: Taking Stylometry to the Underground. In *IEEE S&P*.
[4] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2, 211–236.
[5] Mishari Almishari, Dali Kaafar et al. 2014. Stylometric Linkability of Tweets. In *WPES*.
[6] Mishari Almishari and Gene Tsudik. 2012. Exploring Linkability of User Reviews. In *ESORICS*.
[7] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11.
[8] Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. 2013. Stylometric Analysis for Authorship Attribution on Twitter. In *BDA*.
[9] Patrick Blogamundo: Hall. 2007 (accessed January 8, 2018). *Languages by Average word length.* https://web.archive.org/web/20090823115022/http://blogamundo.net:80/lab/wordlengths (original website inactive).
[10] Pew Research Center. 2018 (accessed January 4, 2019). *Digital News Fact Sheet.* http://www.journalism.org/fact-sheet/digital-news/.
[11] Marco Cristani, Giorgio Roffo et al. 2012. Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging. In *ACM MM*.
[12] Chris Elliott. 2014 (accessed December 14, 2018). *The readers' editor on… pro-Russia trolling below the line on Ukraine stories.* https://gu.com/p/3pvee.
[13] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *ACL*.
[14] Andrew J. Flanagin and Miriam J. Metzger. 2000. Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly* 77, 3.
[15] Elle Hunt. 2017 (accessed December 14, 2018). Facebook purges tens of thousands of fake accounts to combat spam ring. https://gu.com/p/6afbt.
[16] Fredrik Johansson, Lisa Kaati, and Amendra Shrestha. 2013. Detecting Multiple Aliases in Social Media. In *IEEE/ACM ASONAM*.
[17] Fredrik Johansson, Lisa Kaati, and Amendra Shrestha. 2015. Timeprints for Identifying Social Media Users with Multiple Aliases. *Security Informatics* 4, 1.
[18] Srijan Kumar, Justin Cheng et al. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *WWW*.
[19] Huayi Li, Zhiyuan Chen et al. 2015. Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns. In *AAAI ICWSM*.
[20] Todor Mihaylov, Ivan Koychev et al. 2015. Exposing Paid Opinion Manipulation Trolls. In *ACL RANLP*.
[21] Todor Mihaylov, Tsvetomila Mihaylova et al. 2018. The dark side of news community forums: opinion manipulation trolls. *Internet Research* 28, 5.
[22] Todor Mihaylov and Preslav Nakov. 2016. Hunting for Troll Comments in News Community Forums. In *ACL*.
[23] Eni Mustafaraj and Panagiotis Takis Metaxas. 2017. The Fake News Spreading Plague: Was it Preventable?. In *ACM WebSci*.
[24] Arvind Narayanan, Hristo Paskov et al. 2012. On the Feasibility of Internet-Scale Author Identification. In *IEEE S&P*.
[25] Sergey Sanovich. 2018. *Russia: The Origins of Digital Misinformation.*
[26] Roy Schwartz, Oren Tsur et al. 2013. Authorship Attribution of Micro-Messages. In *ACL EMNLP*.
[27] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI*.
[28] Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2013. A Case Study of Sockpuppet Detection in Wikipedia. In *ACS LASM*.
[29] Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60, 3.
[30] Jon Swaine. 2018 (accessed December 14, 2018). Twitter admits far more Russian bots posted on election than it had disclosed. https://gu.com/p/8vp2h.
[31] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380.
[32] Savvas Zannettou, Barry Bradlyn et al. 2019. Characterizing the Use of Images by State-Sponsored Troll Accounts on Twitter. arXiv:1901.05997
[33] Savvas Zannettou, Tristan Caulfield et al. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM IMC*.
[34] Savvas Zannettou, Tristan Caulfield et al. 2019. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *WWW Companion*.
[35] Savvas Zannettou, Tristan Caulfield et al. 2018. Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls. arXiv:1811.03130
[36] Xueling Zheng, Yiu Ming Lai et al. 2011. Sockpuppet Detection in Online Discussion Forums. In *IIH-MSP*.