Sockpuppet Detection in Wikipedia: A Corpus of Real-World Deceptive Writing for Linking Identities

Thamar Solorio, Ragib Hasan, Mainul Mizan
University of Alabama at Birmingham
Birmingham, Alabama
solorio@cis.uab.edu, ragib@cis.uab.edu, mainul@cis.uab.edu

Abstract

This paper describes a corpus of sockpuppet cases from Wikipedia. A sockpuppet is an online user account created with a fake identity for the purpose of covering abusive behavior and/or subverting the editing regulation process. We used a semi-automated method for crawling and curating a dataset of real sockpuppet investigation cases. To the best of our knowledge, this is the first corpus available on real-world deceptive writing. We describe the process for crawling the data and some preliminary results that can be used as baseline for benchmarking research. The dataset has been released under a Creative Commons license from our project website (`http://docsig.cis.uab.edu/tools-and-datasets/`).

Keywords: sockpuppet detection corpus, authorship identification, deceptive language

1. Introduction

According to Wikipedia's policies, each user is supposed to create only one user account. However, Wikipedia does not enforce the one-user one-account rule through technical means. As a result, users are free to create multiple accounts if they want to. When a user creates a secondary account for malicious purposes it is called a sockpuppet. This ease of creating user accounts has led malicious users to create multiple identities and use them for various purposes, ranging from block evasion, false majority opinion claims, and vote stacking.

One of the main applications of the sockpuppet dataset is to develop an automated tool for sockpuppet detection in Wikipedia. Currently, the process for detecting sockpuppets is manual and involves significant experience from the administrators. In many cases, special Wikipedia administrators with IP-address viewing privileges ("checkusers") have to be involved to check user IP addresses of the suspicious accounts. This violates user privacy. Without accessing the IP addresses, the administrators need to depend on their experience in dealing with sockpuppets to manually detect similarities in writing style and behavior. That leaves a lot of room for error. In contrast, an automated tool trained using our sockpuppet dataset can be used to identify the sockpuppets without requiring IP address information or expert administrator knowledge. In practice, the automated tool can be used to assist administrators to more accurately identify malicious sockpuppets.

This data set can support the design of tools for the automated detection of sockpuppets in Wikipedia, but it also has other potential applications. In particular, this corpus can be used by researchers working on authorship attribution problems. The sockpuppet corpus provides a real world data set of short messages from malicious users. The sockpuppet cases involve text from actual users who are intentionally creating multiple identities and actively trying to hide their relationship to the sockpuppet master. Therefore, using this corpus, researchers can evaluate their approaches, or further investigate the writing characteristics on deceptive language in a real life setting. This type of authorship attribution of short text has potential applications in identifying terrorists in web forums, online discussion boards, phone text messages, tweets and other social media interactions where text tends to be short and full of nuances from spontaneous language.

2. Related Work

Authorship analysis has received a great deal of attention in recent years (Stamatatos, 2008). The field has grown from a pure manual stylistic analysis to machine learning approaches that combine stylistic features with richer representations of writing preferences, such as n-grams of syntactic features (Sidorov et al., 2013) and local histograms of character n-grams (Escalante et al., 2011). Recent work started exploring the limits of automated approaches to the problem of authorship analysis by looking at extremely short documents (Layton et al., 2010), very large candidate sets (Koppel et al., 2011), and cross-domain scenarios (Goldstein-Stewart et al., 2008).

Less work has been devoted to authorship analysis on deceptive writing. Some of the exceptions include the work in (Brennan et al., 2012; Novak et al., 2004). The main barrier to study attribution in adversarial scenarios is the lack of suitable data. This is understandable as the nature of the problem makes it difficult to have readily available data where subjects have been intentionally trying to deceive humans. To solve this barrier researchers have turn to the generation of artificial data sets. For instance Novak et al. generated sub aliases from message boards by randomly splitting data from the same alias (Novak et al., 2004). Then they evaluated performance of their method on linking the two sub aliases. The Brennan-Greenstadt adversarial stylometry corpus was collected from volunteers (Brennan et al., 2012). The authors instructed the subjects to submit original writings of an academic nature. Then the subjects were asked to obfuscate their writing style during the creation of a topic specific writing of 500 words. In addition, subjects were also requested to submit an imitation writing excerpt, where they were instructed to imitate the writing of

Cormac McCarthy in *The Road*. Here again, the topic of the imitation writing was controlled by the corpus developers. Both resources are valuable in that they enable researchers to explore attribution approaches, allowing them to show that in adversarial scenarios state of the art approaches will degrade performance. This gap in performance calls for more research in deceptive writing. However, these two data sets still have an artificial flavor to them since the authors were not self motivated and it is not clear whether this will cause major differences in the final stylistic markers of their writings. The sockpuppet corpus we created is a real-world alternative to the study of deceptive writing in social media. The authors were not aware of someone collecting their writings to study attribution, thus this new data set will allow the study of deceptive writing in the wild.

Note that the study of deceptive writing has identified available corpora containing deceptive language samples[1]. However, our goal is to perform attribution in adversarial scenarios, and more specifically, to link non-deceptive writing to deceptive text from the same person. Thus, datasets that can be used in the detection of deceptive language may not be suitable for our task.

3. Sockpuppet Investigations (SPI) in Wikipedia

Wikipedia allows any editor to request investigation of suspected sockpuppetry. The requester needs to include any evidence of the abusive behavior. Typical evidence includes information about the editing patterns related to those accounts, such as the articles, the topics, vandalism patterns, timing of account creation, timing of edits, and voting pattern in disagreements.

Once a case is filed, an administrator will investigate the case. An administrator is an editor with privileges to make account management decisions, such as banning an editor. The administrator performs a behavioral evidence investigation and will try to determine whether the two accounts are related and will then issue a decision confirming or rejecting the sockpuppetry case, or request involvement of a check user. Check users are higher privileged editors, who have access to private information regarding editors and edits, such as the IP address from which an editor has logged in. Check users perform a technical evidence investigation. But as explained in Wikipedia SPI description, these users will be involved in the investigation, if needed, only after strong behavioral evidence has been collected.

When an SPI concludes with a confirmed sock puppetry verdict, the sockpuppet account will be banned indefinitely. The administrators have the discretion to establish bans or to block the main account as well.

The process to resolve SPI described above is time consuming and expensive. The last time we checked the list of current cases, on 10/23/13, there were 30+ unique SPI cases listed for the month of October. This high rate of cases filed in a single month shows the need for a streamlined process to handle SPIs. The data set we created is a first step in this direction.

[1] See for instance proceedings of the 2012 EACL Workshop on Computational Approaches to Deception Detection.

4. Data Collection Process

All the data we collected from Wikipedia is readily available from the Wikipedia website. Wikipedia archives all information related to each sockpuppet case filed, and once a verdict is issued, that too is stored in the archives. However, because of the lack of a standard format in the archives, our process for data collection is semi-automated. The sockpuppet cases we collected were crawled from the following urls:

- `https://en.wikipedia.org/wiki/Wikipedia:Sockpuppet_investigations/SPI/Closed/2009`
- `https://en.wikipedia.org/wiki/Wikipedia:Sockpuppet_investigations/SPI/Closed/2010`
- `http://en.wikipedia.org/w/index.php?title=Wikipedia:Sockpuppet_investigations/Cases/Overview&offset=&limit=500&action=history`

For each case selected for inclusion in our corpus we collect all data from the talk pages of each editor involved in the SPI case. This step is done automatically by crawling the corresponding Wikipedia archives. We only collect data from discussion pages since these are free form discussions among editors that give them more freedom to show their stylistic writing markers. In contrast, the basic namespaces in Wikipedia, and in particular the articles the editors contribute to, have a more restrictive format that can make difficult the identification of editors. Moreover, some of the edits in the main Wikipedia articles include things like reverts, or typo corrections, that are related to the user behavior and not necessarily to editors writing styles. Our main goal to develop this corpus is to support research in deceptive writing, and thus the behavior treats mentioned above fall outside this goal. However, this information could still be crawled at a later stage and be leveraged to perform a persona identification.

The manual process for this task involves retrieving the final decision reached by the investigative administrator or check user. There is no fixed format for recording decisions on SPI cases and therefore parsing the data with regular expressions will not work for most cases. We were required to visit each SPI case and read the discussion of any administrators investigating the case and check users involved. This was the bottle neck for the process and what prevented us from having a larger sample, although we continue to add cases to our data set as feasible.

The majority of the SPI cases in Wikipedia end up being confirmed as sock puppets. This is reasonable since editors file cases after they have already seen some suspicious behavior. Therefore, to provide a larger number of non-sock puppet cases, we crawled pairs of editors that have not been involved in SPI before but that have participated in the same talk pages as editors involved in SPI cases.

Comment from the sockpuppeteer: Inanna Mine was original and i have worked on it more than 4 hours.**I** have changed it many times by opinions.**L**ast one was accepted by all the users(except for khokhoi).**I** have never used sockpuppets.**P**lease dont care Khokhoi,Tombseye and Latinus.They are changing all the articles about Turks.The most important and famous people are on my picture.
Comment from the sockpuppet: Altau Hello.**I** am trying to correct uncited numbers in Battle of Sarikamis and Crimean War by resources but khoikhoi and tombseye always try to revert them.**C**ould you explain them there is no place for hatred and propagandas, please?
Comment from another editor: Khoikhoi Actually, my version WAS the original image. Ask any other user. Inanna's image was uploaded later, and was snuck into the page by Inanna's sockpuppet before the page got protected. The image has been talked about, and people have rejected Inanna's image (see above).

Table 1: Sample excerpt from a single sockpuppet case. We show in boldface some of the stylistic features shared between the sockpuppeteer and the sockpuppet.

Confirmed SPI cases	305
Denied SPI cases	105
Created non-sock puppet cases	213
Average number of comments per case	$\sim$ 180
Average number of comments per editors	$\sim$ 83

Table 2: The sockpuppet data set

5. The Sockpuppet Corpus

We originally collected around 700 cases, but after manual inspection we removed about 80 cases where editors did not have content on the talk pages. These were editors that just made contributions directly to Wikipedia pages but did not engage in any side discussions. The resulting corpus currently has 623 cases where 305 of them were confirmed SPI cases by Wikipedia administrators or check users. The remaining 318 are non-sockpuppet cases that combine 105 SPI cases where the administrators verdict was negative, and 213 cases we created from other editors.

Examples from a couple of cases are shown in Table 1. In that table we show a comment from the editor named Inanna that was accused of being the puppeteer of editor Altau. For comparison purposes we show as well a comment made by another editor, not involved in the SPI case on the same talk pages. A noticeable feature in the table is the omission of a white space after the periods.

The table also shows that the comments resemble what we would see in web forum data. For our corpus we found out that the average length in characters is 529. While texts are short, previous work has carried out author identification from tweets (Layton et al., 2010), and many researchers, ourselves included, have reached reasonable prediction performance on social media data that is very similar to the data of this corpus. Some statistics about this dataset are shown in Table 2.

6. A Machine Learning Approach to Sockpuppet Detection

Earlier this year we did a case study of adapting a standard machine-learning authorship attribution approach to predict sockpuppet cases (Solorio et al., 2013). This preliminary study shows some promising results for this task. But it was based on a smaller set of cases, only 77. These 77 cases are a subset of the editors included in the new version of the corpus.

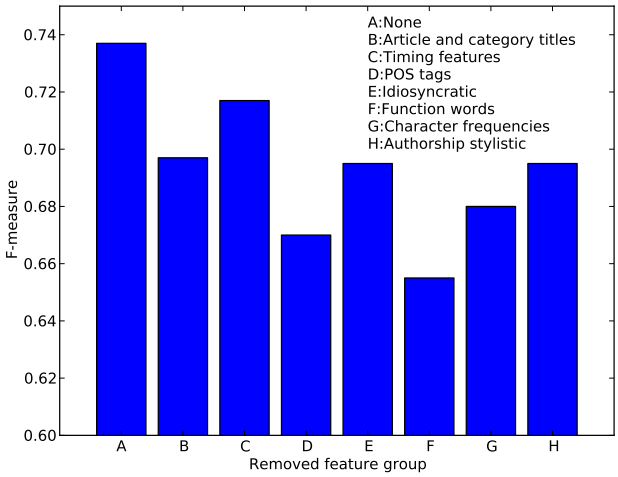

Figure 1: The bars show average F-measures when testing support vector machine removing one feature group at a time in a 10 fold cross-validation setting.

Here we present new results using all 623 cases in a ten-fold cross-validation setting. We hope these results can be used as a sort of baseline comparison for other researchers using this data set.

For these experiments we also changed the underlying framework for the task. Here we assume any pair of editors can be considered an instance of the classification problem, an SPI, and the learner has to decide whether to declare the editors as belonging to the same person or not based on observations from the comments made by each editor involved. The features used in this problem are then the pairwise normalized differences of the feature vectors representing each comment. A complete list of features can be found at the following link: `http://docsig.cis.uab.edu/media/`

2014/03/list-of-features.pdf and a detailed description is in our previous paper (Solorio et al., 2013). Figure 1 shows the results of training a support vector machine (SVM) classifier removing one feature group at a time. We used Weka's implementation of SVMs with default parameters. The best results (F-measure 73%) are achieved using all features. These results are very similar to the results attained in our case study (F-measure 72%).

7. Conclusion

This paper presents a new dataset that will enable research in authorship attribution under real-world adversarial conditions. The nature of the data is very similar to what can be found in social media, which makes it an even more attractive resource as security and privacy concerns in social media data will continue to grow. The prediction results reported here will also be a good baseline for future research. The data set is available from the project website under a Creative Commons license. Our goal is to continue adding SPI cases on a regular basis to maintain an updated resource.

We are currently investigating the use of new features for this problem. Our goal is to engineer features related to the edit history of the users. In particular, we want to extract topic information from the articles the users have contributed to. As future work we want to design an efficient process to identify many to one relationships between editors. It is very common for a sockpuppet master to be related to more than one sockpuppet account.

Acknowledgments

This research was supported in part by ONR grant N00014-12-1-0217 and NSF grants 1350360 and 1254108. The authors would like to thank Abu Awal Md Shoeb for his support in the data curation process.

8. References

Brennan, M., Afroz, S., and Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3):12:1–12:22, Nov.

Escalante, H. J., Solorio, T., and Montes, M. (2011). Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 288–298. Association for Computational Linguistics (ACL).

Goldstein-Stewart, J., Goodwin, K. A., Sabin, R. E., and Winder, R. K. (2008). Creating and using a correlated corpora to glean communicative commonalities. In *Proceedings of LREC 2008*, pages 3029–3035, Marrakech, Morocco, June.

Koppel, M., Schler, J., and Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, 45:83–94.

Layton, R., Watters, P., and Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. In *Second Cybercrime and Trustworthy Computing Workshop, CTC 2010*, pages 1–8, Ballart, VIC, Australia, July.

Novak, J., Raghavan, P., and Tomkins, A. (2004). Anti-aliasing on the web. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 30–39, New York, NY, USA. ACM.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hérnandez, L. (2013). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*.

Solorio, T., Hasan, R., and Mizan, M. (2013). A case study of sockpuppet detection in wikipedia. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 59–68, Atlanta, Georgia, June. Association for Computational Linguistics.

Stamatatos, E. (2008). A survey on modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.