Reading Between the Lines: A Prototype Model for Detecting Twitter Sockpuppet Accounts Using Language-Agnostic Processes

Erin Smith Crabb(✉), Alan Mishler, Susannah Paletz, Brook Hefright, and Ewa Golonka

University of Maryland, College Park, MD, USA
{ecrabb, amishler, paletz, hefright, egolonka}@umd.edu

Abstract. Sockpuppets are online identities controlled by a user or group of users to manipulate the dissemination of information in digital environments. This manipulation can distort computational assessments of public opinion in social media. Using Russian-language Twitter data from the Ukrainian crisis in 2014, we present a proof-of-concept model employing character n-gram methods to detect sockpuppets. Previous research has demonstrated that n-gram authorship attribution methods can capture lexical preferences, including grammatical and orthographic preferences, while also being less computationally intensive than grammatical or compression language models. Additionally, they can be applied to any language data irrespective of orthography. In this study, a Naïve Bayes classifier was constructed using normalized frequencies of parsed character bigrams to contrast author bigram use. The created model illustrated that suspected sockpuppet accounts were less likely to be correctly classified, showing lower precision, recall, and f-measure rates than other accounts, as predicted.

Keywords: Sockpuppetry · Authorship attribution · Character n-grams · Public opinion measurement · Social media

1 Introduction

Sockpuppets are, "fake identities through which members of the Internet community praise or create the illusion of support for the product of one's work, pretending to be a different person," [1]. Although they are not always used to show support, these identities can have profound effects on the information observed and measured in social media, creating the illusion that many more people share or disagree with a viewpoint than actually do. Researchers have previously used authorship attribution methods to try to detect sockpuppet accounts. These methods, including n-gram, grammatical tag, and compression modeling techniques, attempt to identify sockpuppets based on the features of the texts they produce in order to correctly account for their presence in social media analysis.

C. Stephanidis (Ed.): HCII 2015 Posters, Part I, CCIS 528, pp. 656–661, 2015.
DOI: 10.1007/978-3-319-21380-4_111

This preliminary study investigates the sockpuppetry phenomenon, using character n-gram methods to illustrate that a quantitative difference can be observed when comparing genuine authors and suspected sockpuppet accounts. Section 2 contains a brief review of sockpuppetry as a form of persuasion and character n-grams as a method for sockpuppet detection. Section 3 details the methods and dataset we created for this study, and Sect. 4 summarizes our findings. Finally, Sect. 5 offers some suggestions for this research going forward.

2 Sockpuppetry as Persuasion

In psychological theory, persuasion can act either via effortful thinking or via a "peripheral" route, where the persuader and the person persuaded rely on simple, low-effort strategies and heuristics [10, 11]. These heuristics can include repetition of the same argument, source attractiveness (for example, perceived expertise of the source or trust in the source), and emotional manipulation, which may or may not be relevant to the topic being argued. Even before an explicit persuasive appeal is made, the proponent can set the stage by using the pre-persuasion techniques of biased language, framing, and creating or applying norms [12]. The use of sockpuppets draws on pre-persuasion techniques of establishing an anchor for social comparison and presenting a norm. These accounts also rely on repetition, a peripheral persuasion technique, to create the false perception of different sources saying and believing the same thing.

In order to identify sockpuppet accounts, a variety of authorship attribution methods, including n-gram, grammatical, topic modeling, and compression methods, have been applied to computationally contrast authors' texts [1, 3]. Some studies have further complemented these by using non-linguistic features, such as time stamps and profile information [14]. Character n-gram methods involve comparing the frequencies of sequences of n numbers of contiguous characters in documents [2]. (For example, the word Ukraine would include the bigrams ($n = 2$) of *Uk*, *kr*, *ra*, *ai*, *in*, and *ne*.) N-gram methods can be particularly useful for capturing lexical preferences, including grammatical and orthographic preferences [5]. Additionally, unlike grammatical features and function word analysis, character n-grams can be calculated without deep linguistic knowledge of the language being studied, making application of this method much more straightforward and less computationally intensive than methods requiring language-specific knowledge and resources [5].

Luyckx and Daelemans [9] explore the application of character n-gram authorship attribution models to Twitter data, concluding that of all the methods they tried, character n-grams were the most successful at predicting authorship. Kukushkina, Polikarpov, and Khmelev [6], the only study we are aware of which addresses authorship attribution of Russian data, find that character bigrams are the most reliable and accurate features for predicting authorship of Russian-language documents, scoring higher than the frequency of word n-grams, single grammatical classes, and grammatical class bigrams. Because of the ease of application and the previous success of this method with Russian, we utilize the character bigram model to compare authors quantitatively in order to identify sockpuppets.

3 Methods

The dataset for this study was comprised of Twitter messages, or "tweets," obtained in JSON format from TweetTracker, a web-based portal for collecting tweets and other social media [7, 8]. We identified the harvested tweets via the "Crimea" TweetTracker query covering a two-month period from June 25 to August 25, 2014. Utilizing this query to collect tweets ensures that they either include a key word or hashtag referencing a set of pre-defined key words pertaining to Crimea or Ukraine, are a particular user, or are geotagged as originating from the area defined by a geographic coordinate bounding box. Prior to identifying authors for this study, we selected Russian in TweetTracker as the preferred language of returned tweets [7].

From this set of possible tweets, we chose 23 authors who had posted more than one hundred tweets between June 25, 2014 and August 25, 2014, and exported their data. Based on our reading of a sample of the tweets, we suspected that three of the 23 authors were sockpuppets (accounts K, L, and Q: see Table 1); these accounts seemed dedicated to tweeting news headlines, and spot-checking identified some very similar

Table 1. Precision, recall and f-measure scores for each author, followed by weighted averages

Accounts	Precision	Recall	F-measure
A	0.214	0.18	0.196
B	0.747	0.59	0.659
C	0.898	0.97	0.933
D	0.182	0.18	0.181
E	0.374	0.43	0.4
F	0.511	0.7	0.591
G	0.169	0.14	0.153
H	0.234	0.18	0.203
I	0.306	0.26	0.281
J	0.282	0.2	0.234
K	**0.192**	**0.23**	**0.209**
L	**0.089**	**0.11**	**0.099**
M	0.292	0.21	0.244
N	0.145	0.16	0.152
O	0.273	0.41	0.328
P	0.156	0.17	0.163
Q	**0.04**	**0.05**	**0.045**
R	0.189	0.14	0.161
S	0.541	0.4	0.46
T	0.635	0.73	0.679
U	0.237	0.18	0.205
V	0.261	0.35	0.299
W	0.21	0.17	0.188
Weighted Average:	*0.312*	*0.31*	*0.307*

tweets. For each account, we selected the first one hundred tweets and used them to create the n-gram models discussed below. We chose this number of tweets because Luyckx and Daelemans [9] state that their models have had success attributing texts to authors for whom they only have a minimum of 1,400 words, although providing more data could increase the model's accuracy [5, 9]. Thus, by providing one hundred tweets, our models should have far more data than this minimum to work with.

Prior to the analysis, we conducted several pre-processing steps on the 23-author dataset to ensure that the tweets contained only the authors' actual text. First, we used a computational algorithm to delete non-UTF-8 characters, such as the "thumbs-up" and "house" symbols. We then removed hyperlinks, user mentions, and hashtags. These have a high probability of producing uninformative features because of their purely lexical value: because they are so rare, they can result in overfitting for particular authors. Additionally, they may not be the author's own words. Finally, we removed labeled retweets from each author's data.

We then extracted n-gram features from the cleaned tweets, decomposing them into their component character bigrams and calculating the frequency of each bigram at the tweet level to normalize the values. The resulting output was a feature matrix displaying the frequencies of each bigram normalized by the total number of bigrams per tweet.

We used WEKA [4], a tool implementing a variety of classification and clustering algorithms, to analyze the resulting feature matrix. We tried several classifiers, using ten cross-fold validation to test their performance, and obtained the most compelling results with WEKA's implementation of a general Naïve Bayes classifier [4]. In addition to scoring the classifiers, WEKA provided a confusion matrix displaying how each author's tweets were assigned, which we discuss further below [4].

4 Findings

Under the assumption that all 23 authors were distinct, the classifier had a weighted average precision score of 0.312 and a recall score of 0.31, for an f-measure of 0.307. As seen in Table 1, some authors had much higher individual f-measures (as high as f = 0.933) and others had much lower measures (f = 0.045). Low precision and recall rates are expected for authors with very similar or identical content. Thus, rather than interpreting the overall f-measure, precision score, or recall score as an absolute ideal number, the utility of this method is to identify specific authors and examine the hypothesis that the authors with low scores are not all distinct (i.e., are potential sockpuppets).

Notably, two of the lowest score sets belong to accounts L and Q, which were two of the suspected sockpuppet accounts. The third suspected sockpuppet account, K, achieved a higher f-measure, but it was still in the lower half of scores. When we re-ran the classifier without the three suspected sockpuppets, the average f-measure increased to 0.35. Together, these findings indicate that these accounts may very well be sockpuppets, as their tweets were highly confusable with those belonging to other authors.

We obtained additional results from the classifier's confusion matrix, an abbreviated section of which can be seen in Table 2. For 20 of the 23 authors, the largest

Table 2. Abbreviated classifier confusion matrix for B, C, K, L, and Q

	B	C	K	L	Q
B	**59**	2	6	2	3
C	0	**97**	0	3	0
K	0	3	23	25	**31**
L	4	2	24	11	**51**
Q	0	1	32	**56**	5

number of tweets assigned to any of the 23 was identified as belonging to the correct author – that is, for those accounts not initially suspected of being sockpuppets, the largest number of tweets classified as belonging to any given author was correctly associated with that author. In Table 2 below, this can be seen as the 59 and 97 tweets which were accurately assigned to authors B and C respectively, whereas the possible sockpuppets, K, L, and Q were highly confusable with one another. The algorithm categorized the largest subsets of K and L as belonging to Q, while 56 of author Q's tweets were identified as author L's, indicating that these authors have so much similar or identical content as to be confusable.

Overall, this study shows that using character bigrams is a promising potential technique for identifying sockpuppets. As predicted, certain users previously considered to be sockpuppets because of their content showed dramatically low recall and precision scores, and were highly confusable with one another.

5 Future Work

There are several steps to take with this research going forward. It will be important to conduct reliability testing of this method with a ground-truth dataset. To do so, we will use the corpus documented by [13], which has established sockpuppets and genuine users. As Twitter data are stylistically different from forum postings, we will investigate whether or not a synthetic sockpuppet dataset can serve as a viable alternative to a ground-truth dataset by randomly dividing a subset of single-author accounts into multiple "authors," running the classifier, and examining precision and recall values along with confusion matrix results. These findings will be compared with the results obtained from running the algorithm against [13]. Combined, we hope these results will provide an accuracy baseline that would enable us to measure improved techniques for sockpuppet detection. We will also test how increasing authors and linguistic diversity within the data may affect the accuracy results for sockpuppet detection.

References

1. Bu, Z., Xia, Z., Wang, J.: A sockpuppet detection algorithm on virtual spaces. Knowl.-Based Syst. **37**, 366–377 (2013)

2. Cavnar, W., Trenkle, J.: N-gram-based text categorization. In: Proceedings of SDAIR-1994, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175. Information Science Research Institute, Las Vegas (1994)
3. Fornaciari, T., Poesio, M.: Identifying fake Amazon reviews as learning from crowds. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 279–287. Association for Computational Linguistics (2014)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. SIGKDD Explorations **11**(1), 10–18 (2009)
5. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inform. Sci. Technol. **60**(1), 9–26 (2009)
6. Kukushkina, O., Polikarpov, A., Khmelev, D.: Using literal and grammatical statistics for authorship attribution. Probl. Inf. Transm. **37**(2), 172–184 (2001)
7. Kumar, S., Barbier, G., Abbasi, M., Liu, H.: TweetTracker: An analysis tool for humanitarian and disaster relief. In: Proceedings of the International Conference on Weblogs and Social Media, pp. 661–662. The AAAI Press, Palo Alto (2011)
8. Kumar, S., Morstatter, F., Liu, H.: Twitter Data Analytics. Springer, New York (2013)
9. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing **26**(1), 35–55 (2011)
10. Petty, R., Cacioppo, J.: The elaboration likelihood model of persuasion. Adv. Soc. Psychol. **19**, 123–205 (1986)
11. Petty, R., Cacioppo, J., Strathman, A., Priester, J.: To think or not to think: Exploring two routes to persuasion. In: Brook, T.C., Green, M.C. (eds.) Persuasion: Psychological Insights and Perspectives, pp. 81–116. Sage, Thousand Oaks (2005)
12. Pratkanis, A., Aronson, E.: Age of Propaganda: The Everyday Use and Abuse of Persuasion. W. H. Freeman, New York (2001)
13. Solorio, T., Ragib, H., Mizan, M.: Sockpuppet detection in Wikipedia: A corpus of real-world deceptive writing for linking identities. Computing Research Repository (2013). arXIV: 1310.6772 [cs.CL]
14. Tsikerdekis, M., Zeadally, S.: Multiple account identity deception detection in social media using nonverbal behavior. Library and Information Science Faculty Publications, Paper 13 (2014)