Detection of Opinion Spam with Character n-grams

Donato Hernández Fusilier[1,2], Manuel Montes-y-Gómez[3], Paolo Rosso[1], and Rafael Guzmán Cabrera[2]

[1] Natural Language Engineering Lab.,
Universitat Politècnica de València, Spain
[2] División de Ingenierías, Campus Irapuato-Salamanca,
Universidad de Guanajuato, Mexico
[3] Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico
{donato,guzmanc}@ugto.mx, mmontesg@ccc.inaoep.mx
prosso@dsic.upv.es

Abstract. In this paper we consider the detection of opinion spam as a stylistic classification task because, given a particular domain, the deceptive and truthful opinions are similar in content but differ in the way opinions are written (style). Particularly, we propose using character n-grams as features since they have shown to capture lexical content as well as stylistic information. We evaluated our approach on a standard corpus composed of 1600 hotel reviews, considering positive and negative reviews. We compared the results obtained with character n-grams against the ones with word n-grams. Moreover, we evaluated the effectiveness of character n-grams decreasing the training set size in order to simulate real training conditions. The results obtained show that character n-grams are good features for the detection of opinion spam; they seem to be able to capture better than word n-grams the content of deceptive opinions and the writing style of the deceiver. In particular, results show an improvement of 2.3% and 2.1% over the word-based representations in the detection of positive and negative deceptive opinions respectively. Furthermore, character n-grams allow to obtain a good performance also with a very small training corpus. Using only 25% of the training set, a Naïve Bayes classifier showed F_1 values up to 0.80 for both opinion polarities.

Keywords: Opinion spam, deceptive detection, character n-grams, word n-grams.

1 Introduction

With the increasing availability of review sites people rely more than ever on online opinions about products and services for their decision making. These reviews may be positive or negative, that is, in favour or against them. A recent survey found that 87% of people have reinforced their purchase decisions by

A. Gelbukh (Ed.): CICLing 2015, Part II, LNCS 9042, pp. 285–294, 2015.
DOI: 10.1007/978-3-319-18117-2_21

positive online reviews. At the same time, 80% of consumers have also changed their minds about purchases based on negative information they found online[1]. Additionally, there is a special class of reviews, the *deceptive opinions*, which are fictitious opinions that have been deliberately written to sound authentic in order to deceive the consumers. Due to their growing number and potential influence, the automatic detection of opinion spam has emerged as a highly relevant research topic [3,17,8].

Detecting opinion spam is a very challenging problem since opinions expressed on the Web are typically short texts, written by unknown people for very different purposes. Initially, opinion spam was detected by methods that seek for duplicate reviews [9]. It was only after the release of the gold-standard datasets by [16,17], which contain examples of positive and negative deceptive opinion spam, that it was possible to conduct supervised learning and a reliable evaluation of the task. The main conclusion from recent works is that standard text categorization techniques are effective at detecting deception in text. Particularly, best results have been approached using word n-grams together with other stylometric features [4,17].

We consider the detection of opinion spam as a stylistic classification task because, given a particular domain, the deceptive and truthful opinions are similar in content but differ in the way opinions are written (style). Furthermore, based on the fact that character n-grams are able to capture information from content and style, and motivated by their good performance in other tasks such as authorship attribution and polarity classification, we propose in this paper the use of character n-grams for the detection of opinion spam. Concretely, we aim to investigate in depth whether character n-grams are: (i) more appropriate than word n-grams, and (ii) more robust than the word n-grams in scenarios where only few data for training are available. Two are the main experiments we carried out. In the first experiment we considered 1600 hotel reviews. We analysed the classification of positive and negative opinions employing as features character n-grams and word n-grams. The best results were obtained using character n-grams with values for n of 5 and 4 respectively. The second experiment was varying the size of the training corpus in order to demonstrate the robustness of character n-grams as features. The obtained results show that with few samples in the training corpus, it is possible to obtain a a very classification performance, comparable to that obtained by word n-grams when using the complete training set.

The rest of the paper is organized as follows. Section 2 presents the related works on opinion spam detection and the use of character n-grams in other text classification tasks. Section 3 describes the corpus used for experiments as well as their configuration. Section 4 discusses the obtained results. Finally, Section 5 indicates the main contributions of the paper and provides some directions for future work.

[1] How Online Reviews Affect Your Business.
`http://mwpartners.com/positive-online-reviews`. Visited: April 2, 2014.

2 Related Work

The detection of spam on the Web has been mainly approached as a binary classification problem (spam vs. non-spam). It has been traditionally studied in the context of e-mail [2], and Web pages [5,15]. The detection of opinion spam, i.e., the identification of fake reviews that try to deliberately mislead human readers, is just another face of the same problem [18].

Due to the lack of reliable labeled data, most initial works regarding the detection of opinion spam considered unsupervised approaches which relied on meta-information from reviews and reviewers. For example, in [9], the authors proposed detecting opinion spam by identifying duplicate content. In a subsequent paper [10], they focussed on searching for unusual review patterns. In [14], the authors proposed an unsupervised approach for detecting groups of opinion spammers based on criteria such as the number of products that have been target of opinion spam and a high content similarity of their reviews. Similarly, in [20] it is presented a method to detect hotels which are more likely to be involved in spamming.

It was only after the release of the gold-standard datasets by [16,17], which contain examples of positive and negative deceptive opinion spam, that it was possible to conduct supervised learning and a reliable evaluation of the task. [16,13,3,17,7,8] are some examples of works that have approached the detection of opinion spam as a text classification task. In all of them word n-grams (unigrams, uni+bigrams and uni+bi+trigrams) have been employed as features. However, best results have been obtained combining word n-grams with style information. For example, [16] considered information from LIWC (linguistic inquiry and word count)[2], and [4] incorporated syntactic stylometry information in the form of deep syntax features.

In this work, we propose the use of character n-grams for detecting opinion spam. By using this representation our aim is to focus more on the writing style of the deceptive opinions than in their content. That is, our hypothesis is that somehow the writing style of a deceiver is different if compared to the one of honest users. This was also corroborated by Ott in [16].

Character n-grams have been used for email spam detection [11] and sentiment classification [1] with higher effectiveness than using word n-grams. They are also considered the state-of-the-art for authorship attribution [19]. To the best of our knowledge, this work is the first where character n-grams are used for the detection of opinion spam. The results that we will present in Section 4 show that they allow to address the problem more effectively than with word n-grams.

3 Experimental Setup

To test whether character n-grams allow to address the detection of opinion spam more effectively than word n-grams, we used the corpus of 1600 hotel

[2] `www.liwc.net/`

reviews that was facilitated by Ott[3]. These reviews are about 20 hotels of the downtown area of Chicago, where each hotel has 80 reviews, half of them are positive and the other half are negative. Each positive and negative subset is composed of 20 deceptive reviews and 20 truthful reviews. Deceptive opinions were generated using the Amazon Mechanical Turk, whereas (likely) truthful opinions were mined from reviews on TripAdvisor, Expedia, Hotels.com, Orbitz, Priceline, and Yelp.

The following paragraphs show four opinions for the same hotel. These examples are interesting since they show the great complexity of the automatically, and even manually, detection of deceptive opinions. The opinions are similar and just minor details can help distinguishing one from the other. For example, in [16] authors describe that there is a relationship between deceptive language and imaginative writing, and that deceptive reviews tend to use the words "experience", "my husband", "I", "feel", "business", and "vacation" more than genuine ones.

Example of a positive *deceptive* opinion

My husband and I stayed for two nights at the Hilton Chicago, and enjoyed every minute of it! The bedrooms are immaculate, and the linens are very soft. We also appreciated the free WiFi, as we could stay in touch with friends while staying in Chicago. The bathroom was quite spacious, and I loved the smell of the shampoo they provided-not like most hotel shampoos. Their service was amazing, and we absolutely loved the beautiful indoor pool. I would recommend staying here to anyone.

Example of a positive *truthful* opinion

We stay at Hilton for 4 nights last march. It was a pleasant stay. We got a large room with 2 double beds and 2 bathrooms, The TV was Ok, a 27' CRT Flat Screen. The concierge was very friendly when we need. The room was very cleaned when we arrived, we ordered some pizzas from room service and the pizza was ok also. The main Hall is beautiful. The breakfast is charged, 20 dollars, kinda expensive. The internet access (WiFi) is charged, 13 dollars/day. Pros: Low rate price, huge rooms, close to attractions at Loop, close to metro station. Cons: Expensive breakfast, Internet access charged. Tip: When leaving the building, always use the Michigan Ave exit. It's a great view.

Example of a negative *deceptive* opinion

I stayed two nights at the Hilton Chicago. That was the last time I will be staying there. When I arrived, I could not believe that the hotel did not offer free parking. They wanted at least $10. What am I paying for when I stay there for the night? The website also touted the clean linens. The room was clean and I believe the linens were clean. The problem was with all of the down pillows etc. Don't they know that people have allergies? I also later found out that this hotel allows pets. I think that this was another part of my symptoms. If you like a clean hotel without having allergy attacks I suggest you opt for somewhere else to stay. I did not like how they nickel and dimed me in the end for parking. Beware hidden costs. I will try somewhere else in the future. Not worth the money or the sneezing all night.

Example of a negative *truthful* opinion

My $200 Gucci sunglasses were stolen out of my bag on the 16th. I filed a report with the hotel security and am anxious to hear back from them. This was such a disappointment, as we liked the hotel and were having a great time in Chicago. Our room was really nice, with 2 bathrooms. We had 2 double beds and a comfortable hideaway bed. We had a great view of the lake and park. The hotel charged us $25 to check in early (10am).

[3] http://myleott.com/op_spam

For representing the opinion reviews we used a bag of character n-grams (BOC) and a bag of word n-grams (BOW); in both cases we applied a binary weighting scheme. Particularly, for building the BOW representation we preprocessed texts removing all punctuation marks and numerical symbols, i.e., we only considered alphabetic tokens. We maintained stop words, and converted all words to lowercase characters.

For classification we used the Naïve Bayes (NB) classifier, employing the implementation given by Weka [6], and considering as features those n-grams that occurred more than once in the training corpus. It is important to comment that we performed experiments using several classification algorithms (e.g., SVM, KNN and multinomial NB), and from all of them NB consistently showed the best results.

The evaluation of the classification effectiveness was carried out by means of the macro average F_1-measure of the deceptive and truthful opinion spam categories. We performed a 10 fold cross-validation procedure to assess the effectiveness of each approach, and we used the the Wilcoxon Signed Rank Test for comparing the results of BOC and BOW representations in all the evaluation scenarios. For these comparisons we considered a 95% level of significance (i.e., $\alpha = 0.05$) and a null hypothesis that both approaches perform equally well.

4 Experiments

In this section we describe the two experiments we carried out in order to see whether character n-grams allow to obtain a better performance than word n-grams (first experiment), and also to evaluate the robustness of character-based representation when only few examples of deceptive opinion spam are available for training (second experiment).

4.1 Experiment 1: Character vs. Word n-grams

In this first experiment, we aim to demonstrate that character n-grams are more appropriate than word n-grams to represent the content and writing style of opinion spam. We analysed the performance of both representations on positive as well as on negative reviews.

Table 1 shows the results obtained with word n-grams. These results indicate that the combination of unigrams and bigrams obtained the best results in both polarities; however, the difference in F_1 with unigrams was not statistically significant in any case. In contrast, the representation's dimensionality was increased 7.5 for the positive opinions and 8 times for the negative reviews, suggesting that word unigrams are a good representation for this task.

Another interesting observation from Table 1 is that classifying negative opinions is more difficult than classifying positive reviews; the highest F_1 measure obtained for negative opinions was 0.848, whereas for positive opinions the best configuration obtained 0.882. We figure out that this behaviour could be caused by the differences in the vocabularies' sizes; the vocabulary employed in negative opinions was 37% larger than the vocabulary from positives, indicating that

Table 1. Results using *word n-grams* as features, in positive and negative opinions. In each case, the reported results correspond to the macro average F_1 of the deceptive and truthful opinion categories.

FEATURES	POSITIVE		NEGATIVE	
	size	*macro* F_1	*size*	*macro* F_1
unigrams	5920	0.880	8131	0.850
uni+bigrams	44268	0.882	65188	0.854
uni+big+trigrams	115784	0.881	174016	0.840

their content is in general more detailed and diverse, and, therefore, that larger training sets are needed for their adequate modelling.

Figure 1 shows the results obtained with character n-grams for different values of n. It also compares these results against the best result using word n-grams as features. These results indicate that character n-grams allow to obtain better results than word n-grams on the positive opinions. The best result was obtained with 5-grams ($F_1 = 0.902$), indicating an improvement of 2.3% over the result using as features word unigrams and bigrams ($F_1 = 0.882$).

Regarding the negative opinions, results were very similar; character n-grams showed to be better than word n-grams. However, in this case the best results were obtained with character 4-grams. We presume, as before, that this behaviour could be related to the larger vocabulary used in the negative opinions, which make difficult the modelling of large n-grams from the given training set. The best result for character n-grams was $F_1 = 0.872$, indicating an improvement of 2.1% over the result using unigrams and bigrams as features $F_1 = 0.854$.

Using the Wilcoxon test as explained in Section 3, we found that the best results from character n-grams are significantly better that the best results from the word-based representations with $p < 0.05$ in the two polarities.

To have a deep understanding of the effectiveness of character n-grams as features, we analysed the 500 n-grams with the highest information gain for both polarities. From this analysis, we have observed that n-grams describing the location of the hotel (e.g. *block, locat, an ave*) or giving some general information about the rooms (e.g. *hroom, bath, large*) are among the most discriminative for positive spam. In contrast, some of the most discriminative n-grams for negative opinions consider general characteristics (e.g. *luxu, smel, xpen*) or they are related to negative expressions (e.g. *_don, (non, nt_b*). This analysis also showed us that the presence of n-grams containing personal pronouns in first person of singular and plural such as *I, my, we* are 20% more abundant in the list of n-grams from negative opinions than in the list from the positive reviews.

4.2 Experiment 2: Character n-grams Robustness

The second experiment aims to demonstrate the robustness of the character n-grams with respect to the size of the training corpus. To carry out this experiment, for each one of the ten folds used for evaluation, we considered 25%, 50%

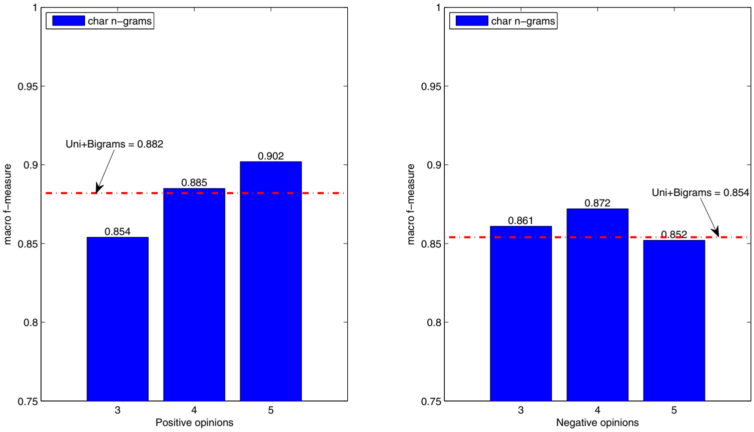

Fig. 1. Results using *character n-grams* as features, in positive and negative opinions. In each case, the reported results correspond to the macro average F_1 of the deceptive and truthful opinion categories. The dotted line indicates the results using word unigrams and bigrams as features.

and 100% of the training instances to train the classifier, while mantaining fixed the test set partition.

Figure 2 shows the results obtained with the Naïve Bayes classifier for both polarities, positive and negative opinions, as well as using both kinds of features, character n-grams and word n-grams. These results indicate that the performance obtained with character n-grams is consistently better that the performance of word n-grams. In particular, it is important to notice that using only 25% of the original training set, which consists of 180 opinions reviews, half of them deceptive and the other half truthful, the representation based on character n-grams shows F_1 values up to 0.80 for both polarities. Using the Wilcoxon test as explained in Section 3, we found that the results from character n-grams are significantly better that the results from the word-based representations with $p < 0.05$ in both polarities.

As an additional experiment we evaluated the variation in performance of the proposed representation using other classifiers. Particularly, Figure 3 compares the results obtained by the Naïve Bayes classifier with those obtained with SVM as well as with a multinomial Naïve Bayes classifier. These results indicate an important variation in F_1 measure caused by the selection of the classifier. On the one hand, the Naïve Bayes classifier shows the best results for the positive opinions; they are significatively better than those from SVM according to the Wilcoxon test with $p < 0.05$. On the other hand, SVM obtained the best results in the classification of deceptive and truthful negative reviews, significantly improving the results of the Naïve Bayes classifier only when using the complete (100%) training set. Somehow this results were not completely unexpected since previous works have showed that Naïve Bayes models tend to surpass the SVM classifiers when there is a shortage of positives or negatives [8].

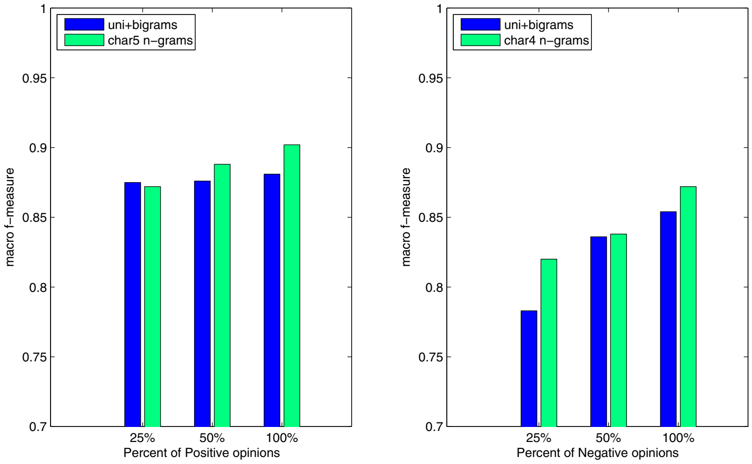

Fig. 2. Results of *character n-grams* and *word n-grams* varying the size of the training sets. The reported results correspond to the macro average F_1 of the deceptive and truthful opinion categories.

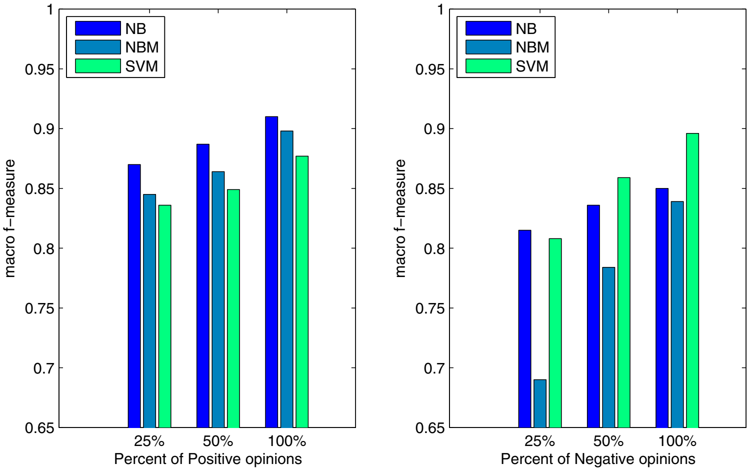

Fig. 3. Results of *character n-grams* using different classifiers and varying the size of the training sets. The reported results correspond to the macro average F_1 of the deceptive and truthful opinion categories.

5 Conclusions and Future Work

In this paper we proposed a novel approach for detecting deceptive opinion spam. We considered the detection of opinion spam as a stylistic classification task, and, accordingly, we proposed using *character n-grams* as features. Although character n-grams have been used in similar tasks showing higher effectiveness that word n-grams, to the best of our knowledge, this work is the first where character n-grams are used for the detection of opinion spam. Experiments were carried out employing Ott's corpus of 1600 hotel reviews, 800 deceptive and 800 truthful. Based on the experimental results it is possible to formulate the following two conclusions: (i) character n-grams showed to capture better than word n-grams the content of deceptive opinions as well as the writing style of deceivers, obtaining better results in both polarities. (ii) Character n-grams showed a better robustness than word-grams obtaining good performance with small training sets; using only 25% of the training data, character n-grams were able to obtained F_1 values up to 0.80 in both polarities.

As future work, we plan to investigate the possibility of combining character n-grams with word n-grams. Going a step forward, we also aim to evaluate other approaches from authorship attribution in the detection of opinion spam.

Acknowledgments. This work is the result of the collaboration in the framework of the WIQEI IRSES project (Grant No. 269180) within the FP 7 Marie Curie. The second author was partially supported by the LACCIR programme under project ID R1212LAC006. Accordingly, the work of the third author was in the framework the DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

1. Blamey, B., Crick, T., Oatley, G.: RU:-) or:-(? character-vs. word-gram feature selection for sentiment classification of OSN corpora. Research and Development in Intelligent Systems XXIX, 207–212 (2012)
2. Drucker, H., Wu, D., Vapnik, V.N.: Support Vector Machines for Spam Categorization. IEEE Transactions on Neural Networks 10(5), 1048–1054 (2002)
3. Feng, S., Banerjee, R., Choi, Y.: Syntactic Stylometry for Deception Detection. Association for Computational Linguistics, short paper. ACL (2012)
4. Feng, S., Xing, L., Gogar, A., Choi, Y.: Distributional Footprints of Deceptive Product Reviews. In: Proceedings of the 2012 International AAAI Conference on WebBlogs and Social Media (June 2012)
5. Gyongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating Web Spam with Trust Rank. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, vol. 30, pp. 576–587. VLDB Endowment (2004)
6. Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: an Update. SIGKDD Explor. Newsl. 10–18 (2009)

7. Hernández-Fusilier, D., Guzmán-Cabrera, R., Montes-y-Gómez, M., Rosso, P.: Using PU-learning to Detect Deceptive Opinion Spam. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, pp. 38–45 (2013)
8. Hernández-Fusilier, D., Montes-y-Gómez, M., Rosso, P., Guzmán-Cabrera, R.: Detecting Positive and Negative Deceptive Opinions using PU-learning. Information Processing & Management (2014), doi:10.1016/j.ipm.2014.11.001
9. Jindal, N., Liu, B.: Opinion Spam and Analysis. In: Proceedings of the International Conference on Web Search and Web Data Mining, pp. 219–230 (2008)
10. Jindal, N., Liu, B., Lim, E.: Finding Unusual Review Patterns Using Unexpected Rules. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010, pp. 210–220(October 2010)
11. Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E.: Word versus character n-grams for anti-spam filtering. International Journal on Artificial Intelligence Tools 16(6), 1047–1067 (2007)
12. Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., Lauw, H.W.: Detecting Product Review Spammers Using Rating Behaviours. In: CIKM, pp. 939–948 (2010)
13. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lecture on Human Language Technologies. Morgan & Claypool Publishers (2012)
14. Mukherjee, A., Liu, B., Wang, J., Glance, N., Jindal, N.: Detecting Group Review Spam. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 93–94 (2011)
15. Ntoulas, A., Najork, M., Manasse, M., Fetterly, D.: Detecting Spam Web Pages through Content Analysis. Transactions on Management Information Systems (TMIS), 83–92 (2006)
16. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding Deceptive Opinion Spam by any Stretch of the Imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 309–319 (2011)
17. Ott, M., Cardie, C., Hancock, J.T.: Negative Deceptive Opinion Spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, pp. 309–319 (2013)
18. Raymond, Y.K., Lau, S.Y., Liao, R., Chi-Wai, K., Kaiquan, X., Yunqing, X., Yuefeng, L.: Text Mining and Probabilistic Modeling for Online Review Spam Detection. ACM Transactions on Management Information Systems 2(4), Article: 25, 1–30 (2011)
19. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law & Policy 21(2) (2013)
20. Wu, G., Greene, D., Cunningham, P.: Merging Multiple Criteria to Identify Suspicious Reviews. In: RecSys 2010, pp. 241–244 (2010)
21. Xie, S., Wang, G., Lin, S., Yu, P.S.: Review Spam Detection via Time Series Pattern Discovery. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 635–636 (2012)
22. Zhou, L., Sh, Y., Zhang, D.: A Statistical Language Modeling Approach to Online Deception Detection. IEEE Transactions on Knowledge and Data Engineering 20(8), 1077–1081 (2008)