Classification of deceptive opinions using a low dimensionality representation

Leticia C. Cagnina
LIDIC
Universidad Nacional de San Luis
San Luis, Argentina
lcagnina@unsl.edu.ar

Paolo Rosso
NLE Lab, PRHLT Research Center
Universitat Politècnica de València
Valencia, España
prosso@dsic.upv.es

Abstract

Opinions in social media play such an important role for customers and companies that there is a growing tendency to post fake reviews in order to change purchase decisions and opinions. In this paper we propose the use of different features for a low dimension representation of opinions. We evaluate our proposal incorporating the features to a Support Vector Machines classifier and we use an available corpus with reviews of hotels in Chicago. We perform comparisons with previous works and we conclude that using our proposed features it is possible to obtain competitive results with a small amount of features for representing the data. Finally, we also investigate if the use of emotions can help to discriminate between truthful and deceptive opinions as previous works show to happen for deception detection in text in general.

1 Introduction

Spam is commonly present on the Web through of fake opinions, untrue reviews, malicious comments or unwanted texts posted in electronic commerce sites and blogs. The purpose of those kinds of spam is promote products and services, or simply damage their reputation. A *deceptive* opinion spam can be defined as a fictitious opinion written with the intention to sound authentic in order to mislead the reader. An opinion spam usually is a short text written by an unknown author using a not very well defined style. These characteristics make the problem of automatic detection of opinion spam a very challenging problem.

First attempts for solving this problem considered unsupervised approaches trying to identify duplicate content (Jindal and Liu, 2008), and searching for unusual review patterns (Jindal et al., 2010) or groups of opinion spammers (Mukherjee et al., 2011). Later, supervised methods were presented. Such is the case of (Feng et al., 2012a; Feng et al., 2012b) in which the authors extended the n-gram feature by incorporating syntactic production rules derived from probabilistic context free grammar parse trees. In (Liu et al., 2002) a learning from positive and unlabeled examples (PU-learning) approach was successfully applied to detect deceptive opinion spam, using only few examples of deceptive opinions and a set of unlabeled data. Then, in (Hernández Fusilier et al., 2015a) the authors proposed a PU-learning variant for the same task, concluding the appropriateness of their approach for detecting opinion spam.

In this paper we study the feasibility of the application of different features for representing safely information about clues related to fake reviews. We focus our study in a variant of the stylistic feature character n-grams named character n-grams in tokens. We also study an emotion-based feature and a linguistic processes feature based on LIWC variables. We evaluated the proposed features with a Support Vector Machines (SVM) classifier using a corpus of 1600 reviews of hotels (Ott et al., 2011; Ott et al., 2013). We show an experimental study evaluating the single features and combining them with the intention to obtain better features. After that previous study, we selected the one with we obtained the best results and made direct and indirect comparisons with some other methods. The obtained results show that the proposed features can capture information from the contents of the reviews and the writing style allowing to obtain classification results as good as with traditional character n-grams but with a lower dimensionality of representation.

The rest of the paper is organized as follows. Section 2 describes briefly the proposed features. Section 3 shows the experimental study performed.

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 58–66,
Lisboa, Portugal, 17 September, 2015.

The description of the corpus and the different experiments carried out can also be found in this section. Finally, the main conclusions and future work are in Section 4.

2 Feature Selection for Deception Clues

In this section we describe the three different kinds of features studied in this work and the tools used for their extraction.

2.1 Character n-grams in tokens

The main difference of character n-grams in tokens[1] with respect to the traditional NLP feature character n-grams is the consideration of the tokens for the extraction of the feature. That is, tokens with less than n characters are not considered in the process of extraction neither blank spaces. Character n-grams in tokens preserve the main characteristics of the standard character n-grams (Šilić et al., 2007): *effectiveness* for quantifying the writing style used in a text (Keselj et al., 2003; Stamatatos, 2013), the *independence* of language and domains (Wei et al., 2008), the *robustness* to noise present in the text (Cavnar and Trenkle, 1994), and, *easiness* of extraction in any text. But unlike the traditional character n-grams, the proposed feature obtains a smaller set of attributes, that is, character n-grams in tokens avoids the need of feature dimension reduction. Figure 1 illustrates that difference.

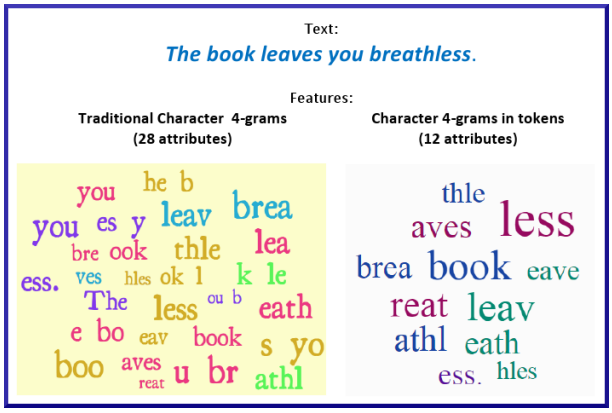

Figure 1: Set of attributes obtained with traditional character n-grams and character n-grams in tokens, considering n=4.

As it can be observed from Figure 1 the amount of attributes obtained with the character n-grams in tokens feature is considerably less, although the effectiveness of this representation still being good, as we will see in Section 3.
For the extraction of character n-grams in tokens we have used Natural Language Toolkit (NLTK) package (Bird et al., 2009) with Python language.

2.2 Emotions-based feature

Previous works have been demonstrated that the use of emotions helps to discriminate truthful from deceptive text (Hancock et al., 2008; Burgoon et al., 2003; Newman et al., 2003). There is some evidence that liars use more negative emotions that truth-tellers. Based on that, we obtained the percentages of positive, negative and neutral emotions contained in the sentences of a document. Then, we have used these values as features in order to represent the polarity of the text.
For the calculation of the percentages of positive, negative and neutral emotions contained in the text we have used the Natural Language Sentiment Analysis API[2] which analyzes the sentiments, labeling a text with its polarity (positive, negative or neutral). We have obtained the polarities of each sentence and then we have obtained the percentages of the polarities associated to the whole document (a review in our case). Finally, we have used those values as features.

2.3 LIWC-based feature: linguistic processes

Several features derived from *Linguistic Inquiry and Word Count* (LIWC) were considered. In particular we have studied those related to functional aspects of the text such as word count, adverbs, pronouns, etc. After performing an early experimental study considering the 26 different variables of the linguistic processes category in LIWC2007 software (Pennebaker et al., 2007), we have concluded that pronouns, articles and verbs (present, past and future tense) would help to distinguish fake from true reviews.

3 Experimental Study

In order to evaluate our proposal, we have performed some experimental study on the first publicly available opinion spam dataset gathered and presented in (Ott et al., 2011; Ott et al., 2013). We first describe the corpus and then we show the different experiments made. Finally we compare our results with those published previously.

[1] Token is considered in this works as any sequence of consecutive characters separated by one or more blank spaces.

[2] http://text-processing.com/demo/sentiment/

3.1 Opinion Spam corpus

The Opinion Spam corpus presented in (Ott et al., 2011; Ott et al., 2013) is composed of 1600 *positive* and *negative* opinions for hotels with the corresponding gold-standard. From the 800 *positive* reviews (Ott et al., 2011), the 400 truthful where mined from TripAdvisor 5-star reviews about the 20 most popular hotels in Chicago area. All reviews were written in English, have at least 150 characters and correspond to users who had posted opinions previously on TripAdvisor (non first-time authors). The 400 deceptive opinions correspond to the same 20 hotels and were gathered using Amazon Mechanical Turk crowdsourcing service. From the 800 *negative* reviews (Ott et al., 2013), the 400 truthful where mined from TripAdvisor, Expedia, Hotels.com, Orbitz, Priceline and Yelp. The reviews are 1 or 2-star category and are about the same 20 hotels in Chicago. The 400 deceptive reviews correspond to the same 20 hotels and were obtained using Amazon Mechanical Turk.

3.2 Truthful from deceptive opinion classification

We have obtained the representations of the opinion reviews considering the features described in Section 2. For all, we have used term frequency-inverse document frequency (tf-idf) weighting scheme. The only text preprocessing made was convert all words to lowercase characters. Naïve Bayes and SVM algorithms in Weka (Hall et al., 2009) were used to perform the classification. We only show the results obtained with SVM because its performance was the best. For all experiments we have performed a 10 fold cross-validation procedure in order to study the effectiveness of the SVM classifier with the different representations. For simplicity, we have used LibSVM[3] which implements a C-SVC version of SVM with a radial basis function. We have run the classifier with the default parameters. The values reported in the tables correspond to the macro average F-measure as it is reported in Weka.
Tables 1, 2 and 3 show the F-measure obtained with each feature proposed for the Opinion Spam corpus.

Table 1 considers only the positive reviews (800 documents). In the first part of the table, we can observe the F-measure obtained with the single features 3 and 4 grams in tokens and, articles, pronouns and verbs extracted from LIWC2007 (referenced as LIWC for simplicity). With the single emotions-based feature (POSNEG in the table) we did not obtain good results; for that reason these are not included in the first part of the table. In the second part of the table, the combination of each single feature was used as representation of the reviews. The best value is in boldface. As we can observe, the best result (F-measure = **0.89**) was obtained with the combination of 4-grams in tokens and the articles, pronouns and verbs (LIWC). With the combination of 3-grams and LIWC feature the F-measure is quite similar.

Feature	F-measure
3-grams in tokens	0.821
4-grams in tokens	0.871
LIWC	0.697
3 + 4-grams in tokens	0.873
3-grams + POSNEG	0.871
4-grams + POSNEG	0.873
3 + 4-grams + POSNEG	0.877
3-grams + LIWC	0.883
4-grams + LIWC	**0.89**

Table 1: Deceptive opinions detection with SVM for positive reviews of Opinion Spam corpus (800 opinions).

Feature	F-measure
3-grams in tokens	0.826
4-grams in tokens	0.851
LIWC	0.69
3 + 4-grams in tokens	0.832
3-grams + POSNEG	0.827
4-grams + POSNEG	0.851
3 + 4-grams + POSNEG	0.827
3-grams + LIWC	0.85
4-grams + LIWC	**0.865**

Table 2: Deceptive opinions detection with SVM for negative reviews of Opinion Spam corpus (800 opinions).

Table 2 shows the results obtained considering only the negative reviews (800 documents). The best result (F-measure = **0.865**) was obtained with the feature 4-grams in tokens plus LIWC variables. It is interesting to note that similar

[3] https://www.csie.ntu.edu.tw/ cjlin/libsvm/

results (although sightly lower) were obtained also with three more features: the single 4-grams in tokens, the combination of the last one with positive and negative emotions percentages, and also with 3-grams combined with LIWC's tokens.

Feature	F-measure
3-grams in tokens	0.766
4-grams in tokens	0.867
LIWC	0.676
3 + 4-grams in tokens	0.854
3-grams + POSNEG	0.858
4-grams + POSNEG	0.87
3 + 4-grams + POSNEG	0.851
3-grams + LIWC	0.866
4-grams + LIWC	**0.879**

Table 3: Deceptive opinions detection with SVM for positive and negative reviews of Opinion Spam corpus (1600 opinions).

Table 3 shows the classification results considering the whole corpus, that is, the combined case of positive plus negative reviews (1600 documents). The best F-measure (**0.879**) was obtained, as the same as the previous cases, with 4-grams in tokens plus LIWC feature. It is worth noting that with the combination of 4-grams in tokens with POSNEG feature seems to be effective when positive and negative polarities are considered together in deception detection, a fact that is not present when just one polarity is considered (see Tables 1 and 2).

As we can observe from Tables 1, 2 and 3, the differences of F-measure values are quite small. In fact, for the almost similar values like, for example, 4-grams in tokens + LIWC compared with 3-grams + LIWC or 3 + 4-grams + POSNEG (see Table 1) the differences are not statistically significant. Consequently we have selected the one with highest F-measure value (4-grams in tokens + LIWC) for simplicity, but some of the other representations can be used instead. In order to analyze the set of attributes corresponding to the feature 4-grams in tokens combined with LIWC, we have calculated the Information Gain ranking.
From this analysis we have observed that the set of attributes with highest information gain are similar for negative and both together polarities corpora. The study shows that 4-grams in tokens are in the top positions of the ranking and those reveal information related to places (*chic*, *chig*, *igan* for Chicago and Michigan cities), amenities (*floo*, *elev*, *room* for floor, elevator, room) and their characterization (*luxu*, *smel*, *tiny* for luxury, smells and tiny). From the 7th position of the ranking we can observe the first LIWC attributes: pronouns (*my*, *I*, *we*) and after 15th position we can observe verbs (*is*, *open*, *seemed*). Interestingly, the articles can be observed from position 68th in the ranking (*a*, *the*).
Regarding the corpus considering only the positive reviews, the ranking is similar to the cases analyzed before with exception of the pronouns which appear at 1st position (*my*) and at 16th position (*I*, *you*). This fact could indicate the presence of many opinions concerned with their own experience (good) making the personal pronouns one of the most discriminative attribute for positive polarity spam opinion detection. With respect to the characterization of the amenities, the adjectives observed in 4-grams in tokens have to do with positive opinions about those (*elax*, *amaz*, *good* for relax, amazing and good). Figure 2 illustrates the first positions of the ranking of attributes obtained for positive reviews.

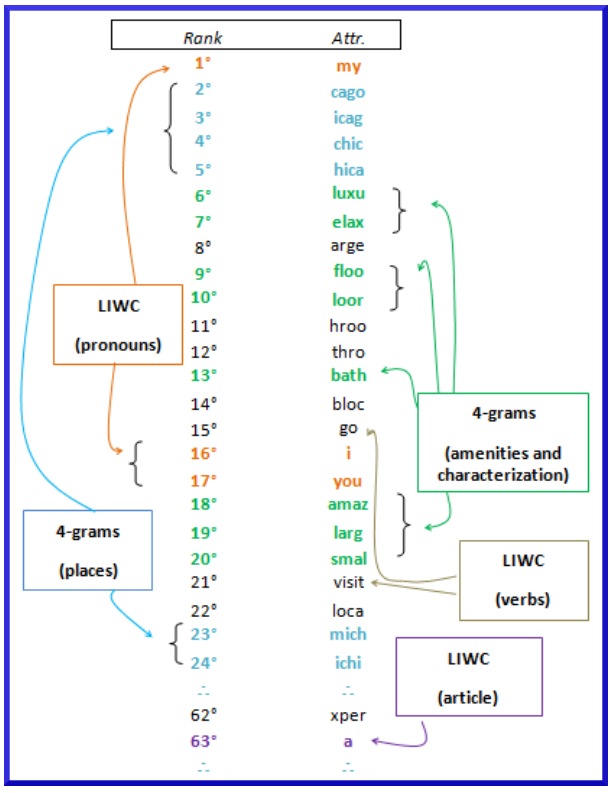

Figure 2: Information gain ranking (partial) for positive reviews.

3.3 Comparison of results

For a comparison of the performance of our proposal, we analyzed the obtained results with respect to the state-of-the-art. We have made a comparison considering the results of five different models. The first four of these were used in an indirect comparison, while just one method was used in a direct comparison of the performance. In (Banerjee and Chua, 2014) the authors presented the results of a logistic regression model using 13 different independent variables: complexity, reading difficulty, adjective, article, noun, preposition, adverb, verb, pronoun, personal pronoun, positive cues, perceptual words and future tense. In (Ren et al., 2014) a semi-supervised model called mixing population and individual property PU learning, is presented. The model is then incorporated to a SVM classifier. In (Ott et al., 2011) the authors used the 80 dimensions of LIWC2007, unigrams and bigrams as set of features with a SVM classifier. In (Feng and Hirst, 2013), profile alignment compatibility features combined with unigrams, bigrams and syntactic production rules were proposed for representing the opinion spam corpus. Then, a multivariate performance measures version of SVM classifier (named SVM^{perf}) was trained. In (Hernández Fusilier et al., 2015b) the authors studied two different representations: character n-grams and word n-grams. In particular, the best results were obtained with a Naïve Bayes classifier using character 4 and 5 grams as features.

As we stated before, two kinds of comparisons are shown: an indirect (we could not obtain the complete set of results reported by the authors) and a direct (the authors kindly made available the results and a statistical comparison can be performed).
In Table 4 we can observe the indirect comparison of our results with those of (Banerjee and Chua, 2014) and (Ren et al., 2014) obtained with a 10 fold cross validation experiment, and then, with a 5 fold cross validation in order to make a fair comparison with the results of (Ott et al., 2011) and (Feng and Hirst, 2013). Note that the results are expressed in terms of the accuracy as those were published by the authors; the results correspond only to positive reviews of the Opinion Spam corpus because the authors experimented in that corpus alone.

From the Table 4 we can observe that the combination of 13 independent variables seems to have the lowest prediction accuracy (accuracy = 70.50%). About the last result, the authors in (Banerjee and Chua, 2014) concluded that only articles and pronouns (over the 13 variables) could significantly distinguish true from false reviews. The accuracy of the semi-supervised model is slightly lower (86.69%) than that of our approach (**89%**), although good enough. The authors concluded that the good performance of the semi-supervised model is due the topic information captured by the model combined with the examples and their similarity (Ren et al., 2014). Then, they could obtain an accurate SVM classifier. Regarding the experiments with the 5 fold cross-validation, we obtained similar results to those of (Ott et al., 2011) and slightly lower than the ones of (Feng and Hirst, 2013). From this last experiment we can observe that using the representation of (Feng and Hirst, 2013) with more than 20138 attributes it is possible to obtain comparable results with those of our approach where we use a smaller representation (1533 attributes).

Model	**Accuracy**
10 fold cross-validation	
(Banerjee and Chua, 2014)	70.50%
(Ren et al., 2014)	86.69%
Our approach	**89%**
5 fold cross-validation	
(Ott et al., 2011)	89.8%
(Feng and Hirst, 2013)	**91.3**%
Our approach	89.8%

Table 4: Indirect comparison of the performance. Deceptive opinions detection for positive reviews of Opinion Spam corpus (800 opinions).

In Table 5 we can observe the direct comparison of the performance for the positive and negative polarities reviews of the Opinion Spam corpus considering the proposal of (Hernández Fusilier et al., 2015b). First column shows the representation proposed, the second one shows the amount of attributes (Attr.) of the representation, the third column shows the F-measure value (F) obtained after a 10 fold cross-validation process, and the last column shows the p-value obtained in the statistical significance test used to study the differences of performance between (Hernández Fusilier et al., 2015b) approach and ours.

Positive reviews (800 opinions)

Model	Attr.	F	p-value
Character 5-grams*	60797	0.90	0.094
Our approach	1533	0.89	

Negative reviews (800 opinions)

Model	Attr.	F	p-value
Character 4-grams*	32063	0.872	0.748
Our approach	1497	0.865	

* (Hernández Fusilier et al., 2015b).

Table 5: Direct comparison of the performance for deceptive opinions detection.

It is interesting to note that the F-measure values obtained with both approaches are quite similar for positive and negative reviews, as we can observe in Table 5. Regarding the amount of attributes used for each representation of the reviews, it is worth noting that our approach uses 97% and 95% fewer attributes for positive and negative reviews compared with the model of (Hernández Fusilier et al., 2015b). Even using a combination of two simple features as character 4-grams in tokens and LIWC variables as we have proposed, the amount of attributes is considerably lower than the traditional character n-grams without diminishing the quality of the classification. The reason of the lower dimensionality of our representation has to do with the manner in which the n-grams are obtained. The high descriptive power of character n-grams in tokens plus the information added with the LIWC variables seem to be adequate to obtain an accurate classifier (SVM in our case).

In order to determine if the differences of performance of (Hernández Fusilier et al., 2015b) and our approach are statistically significant, we have calculated the Mann-Whitney U-test (Mann and Whitney, 1947). This nonparametric test compares two unpaired groups of values without making the assumption of the normality of the samples. However, the requirements of independence of the samples, the data is continuous and ordinal, there are no ties between the groups and the assumption that the distribution of both groups are similar in shape, are satisfied. The null hypothesis states that the samples come from the same population, that is, the classifiers performs equally well with the proposed models. We have calculated the Mann-Whitney U-test considering a 2-tailed hypothesis and significance level of 0.05. In Table 5 we can observe that the p-value obtained in the comparison of performance of positive reviews corpus is $0.094 > 0.05$ which stands for the difference of results are not statistically significant (the p-value is not ≤ 0.05, then the null hypothesis is not rejected). The same conclusion can be obtained with respect to the results corresponding to the negative reviews corpus, for which the test obtained a p-value of $0.748 > 0.05$. From the last test we concluded that both approaches performs similarly well.

A statistical analysis of variance over the F-measure values obtained in the evaluation of (Hernández Fusilier et al., 2015b) and our approach complements our performance study. This analysis can be obtained from the boxplots[4] with the distribution of F-measure values of each proposal with both polarity reviews corpora. Figures 3 and 4 illustrate this analysis. In both figures we can observe that our approach shows a higher dispersion of values, as well as the best F-measure values (0.94 for positive reviews corpus and 0.915 for negative reviews) and the minimum F-measure values (0.84 and 0.81 for positive and negative polarities respectively) compared to the values obtained with (Hernández Fusilier et al., 2015b) approach. However, the median values obtained with both models are quite similar, reason for what there is not statistical difference of performance as it was demonstrated with the statistical significance test.

4 Conclusions and future work

In this work we have proposed some interesting features for deceptive opinions detection. We have studied how different features contribute to model deception clues. Character n-grams in tokens seems to capture correctly the content and the writing style of the reviews helping this, in some way, to differentiate truthful from deceptive opinions. Many works have demonstrated that emotions-based features can discriminate deceptive text, but in our experimental study this feature seems not to provide too much useful information for detecting deception in reviews. We also have used some variables extracted from LIWC

[4]Boxplots (Tukey, 1977) are descriptive statistical tools for displaying information (dispersion, quartiles, median, etc.) among populations of numerical data, without any assumptions about the underlying statistical distribution of the data.

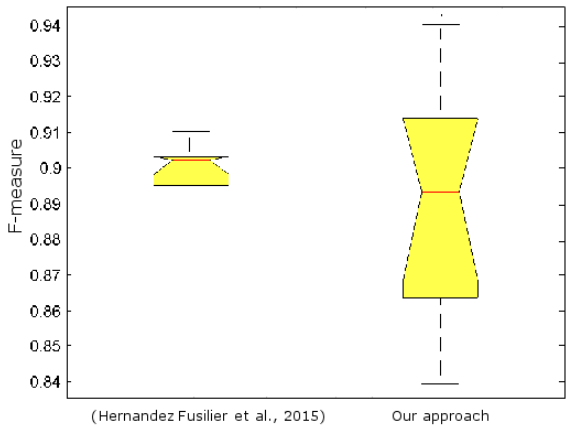

Figure 3: Boxplot for positive reviews corpus in the performance direct comparison.

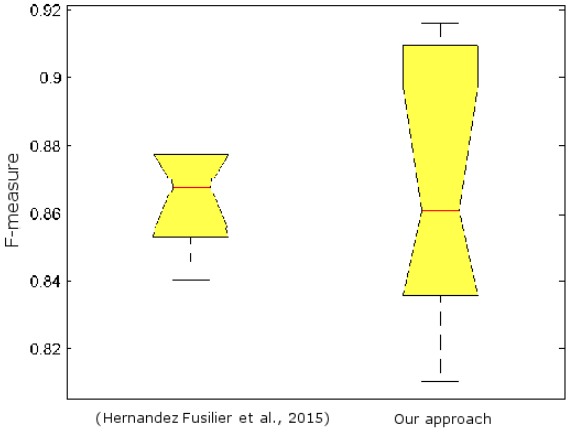

Figure 4: Boxplot for negative reviews corpus in the performance direct comparison.

as pronouns, articles and verbs. That information combined with character 4-grams in tokens was selected for modeling the representation of the reviews. For the experimental study we have used the positive and negative polarities reviews corresponding to the corpora proposed by (Ott et al., 2011; Ott et al., 2013) with 800 reviews each one (400 true and 400 false opinions). We have used both corpora in a separate way but we have performed experiments joining both polarities reviews in a combined corpus of 1600 reviews. From the results obtained with the different features we have concluded that character 4-grams in tokens with LIWC variables performs the best using a SVM classifier. We made also a comparison with the approach of (Hernández Fusilier et al., 2015b) and the results were similar (no statistically significant difference was found), but our low dimensionality representation makes our approach more efficient. For future work we plans to investigate another emotion/sentiment features in order to study the contributions in tasks of deception detection of opinion spam. Also we are interesting to test our model with other corpora related to opinion spam as the one recently proposed in (Fornaciari and Poesio, 2014).

Acknowledgments

The research work of the first author has been partially funded by CONICET (Argentina). The work of the second author was done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems, DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) research project, and the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework on Web Information Quality Evaluation Initiative. The first author would like to thank Donato Hernández Fusilier for kindly providing the results of his method.

References

S. Banerjee and A. Y. K. Chua. 2014. Dissecting genuine and deceptive kudos: The case of online hotel reviews. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Special Issue on Extended Papers from Science and Information Conference 2014:28–35.

S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

J. K. Burgoon, J. P. Blair, T. Qin, and J. F. Nunamaker Jr. 2003. Detecting deception through linguistic analysis. In H. Chen, R. Miranda, D. D. Zeng, C. Demchak, J. Schroeder, and T. Madhusudan, editors, *Intelligence and Security Informatics*, volume 2665 of *Lecture Notes in Computer Science*, pages 91–101. Springer Berlin Heidelberg.

W. B. Cavnar and J. M. Trenkle. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

V. W. Feng and G. Hirst. 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of the 6th International Joint Conference on Natural Language Processing, Nagoya, Japan*, pages 338–346.

S. Feng, R. Banerjee, and Y. Choi. 2012a. Syntactic stylometry for deception detection. In *ACL '12, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 171–175. The Association for Computer Linguistics.

S. Feng, L. Xing, A. Gogar, and Y. Choi. 2012b. Distributional footprints of deceptive product reviews. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 98–105. The AAAI Press.

T. Fornaciari and M. Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287. Association for Computational Linguistics.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth. 2008. On lying and being lied to: a linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.

D. Hernández Fusilier, M. Montes-y-Gómez, P. Rosso, and R. Guzmán Cabrera. 2015a. Detecting positive and negative deceptive opinions using pu-learning. *Information Processing & Management*, 51(4):433 – 443.

D. Hernández Fusilier, M. Montes-y-Gómez, P. Rosso, and R. Guzmán Cabrera. 2015b. Detection of opinion spam with character n-grams. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 9042 of *Lecture Notes in Computer Science*, pages 285–294. Springer International Publishing.

N. Jindal and B. Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230. ACM.

N. Jindal, B. Liu, and E. Lim. 2010. Finding unusual review patterns using unexpected rules. In J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, and A. An, editors, *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1549–1552. ACM.

V. Keselj, F. Peng, N. Cercone, and C. Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, pages 255–264, Dalhousie University, Halifax, Nova Scotia, Canada.

B. Liu, W. S. Lee, P. S. Yu, and X. Li. 2002. Partially supervised classification of text documents. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 387–394. Morgan Kaufmann Publishers Inc.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal. 2011. Detecting group review spam. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 93–94. ACM.

M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. 2003. Lying words: predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1:309–319.

M. Ott, C. Cardie, and J. T. Hancock. 2013. Negative deceptive opinion spam. In *NAACL-HLT 2013, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501. The Association for Computational Linguistics.

J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. 2007. The development and psychometric properties of LIWC2007. In *LIWC webpage. http://www.liwc.net/LIWC2007LanguageManual.pdf*, pages 1–22, Austin, Texas, USA. LIWC.net.

Y. Ren, D. Ji, and H. Zhang. 2014. Positive unlabeled learning for deceptive reviews detection. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 488–498. ACL.

E. Stamatatos. 2013. On the Robustness of Authorship Attribution Based on Character n-gram Features. *Journal of Law and Policy*, 21(2):421–439.

J. W. Tukey. 1977. *Exploratory data analysis*. Pearson Education, Inc., Massachussets, USA.

A. Šilić, J. Chauchat, B. Dalbelo Bašić, and A. Morin. 2007. N-grams and morphological normalization in text classification: A comparison on a croatian-english parallel corpus. In J. Neves, M. F. Santos, and J. M. Machado, editors, *Progress in Artificial Intelligence*, volume 4874 of *Lecture Notes in Computer Science*, pages 671–682. Springer Berlin Heidelberg.

Z. Wei, D. Miao, J. Chauchat, and C. Zhong. 2008. Feature selection on chinese text classification using character n-grams. In *3rd International Conference on Rough Sets and Knowledge Technology (RSKT 08), Chengdu, China*, Lecture Notes in Computer Science, pages 500–507. Springer, Heidelberg, Germany.