Character N-Grams for Detecting Deceptive Controversial Opinions

Javier Sánchez-Junquera[1(✉)], Luis Villaseñor-Pineda[1,3], Manuel Montes-y-Gómez[1], and Paolo Rosso[2]

[1] Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica, San Andres Cholula, Mexico
jjsjunquera@gmail.com, {villasen,mmontesg}@inaoep.mx
[2] PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain
prosso@dsic.upv.es
[3] Centre de Recherche en Linguistique Française GRAMMATICA (EA 4521), Université d'Artois, Arras, France

Abstract. Controversial topics are present in the everyday life, and opinions about them can be either truthful or deceptive. Deceptive opinions are emitted to mislead other people in order to gain some advantage. In the most of the cases humans cannot detect whether the opinion is deceptive or truthful, however, computational approaches have been used successfully for this purpose. In this work, we evaluate a representation based on character n-grams features for detecting deceptive opinions. We consider opinions on the following: abortion, death penalty and personal feelings about the best friend; three domains studied in the state of the art. We found character n-grams effective for detecting deception in these controversial domains, even more than using psycholinguistic features. Our results indicate that this representation is able to capture relevant information about style and content useful for this task. This fact allows us to conclude that the proposed one is a competitive text representation with a good trade-off between simplicity and performance.

Keywords: Deception detection · Controversial opinions
char n-gram

1 Introduction

An opinion is a belief that a person has formed about a topic or issue. It may be about, ideas, laws or experiences, and can be stated with informative purpose or commenting one's own belief about a controversial issue (e.g., politics, health, education, sex, etc.). People need to be provided with significant opinions on current important issues, to form a personal judgment that can impact their future decisions. In order to gain some advantage, there are dishonest opinions whose aim is to mislead thousands of people. Despite the importance of detecting deceptive opinions, psychologists and computational works have proven that it is a very difficult task even for human judges.

Controversial topics are intensively debated in digital media, with opinions expressed in a variety of online forums, newspaper articles, blogs and comments by readers. Even if humans could detect deception effectively, it is inconceivable to manually ensure the authenticity of such opinions.

In previous works, some computational findings are in contrast to psycholinguistic studies of deception. For instance, in [8] the authors observed a lack of negative emotion terms in online hotel reviews, while authors of [2,11] associated negative emotions (e.g. guilt and fear) with interpersonal deception. Thus, cues of deception not only depend on the domain, but also depend on the emotions that a deceiver has about deceiving and its consequences.

Domains of very different nature are, for example, opinion spam and controversial opinions. On hotel reviews, deceivers probably do not have a real opinion about the hotel, so they would be far from having an internal struggle derived from their deception. On the other hand, for a person with an opinion formed about death penalty, to lie about their real point of view is to be against their beliefs, ideals, ethics or religion. In the first case the arguments would be more concrete than in the second, in which the opinion is more philosophical. We presume it is more difficult to detect deception in controversial opinions, [7,8].

Several works have evaluated different text representations for detecting deception in controversial opinions. In [7], some datasets were collected with opinions on abortion, death penalty and personal feelings about the best friend. Using words as features the authors showed that truthful and lying texts are separable. On the same datasets, other approaches have been used with more complex representations such as a combination of words with deep syntactic patterns [4] and with features obtained through Latent Dirichlet Allocation (LDA)[6], both combinations showed good results. Although embeddings are effective in semantically characterizing texts, they do not facilitate the explanation of results, and consequently they have not been popular in this task.

Recently, some character n-grams approaches have been tested for detecting opinion spam. For example, [1,3,5] achieved a very good performance using character n-grams as text representation. Despite several works have employed complex features for the same task in controversial opinions, character n-grams, which are extremely simple features, have not been evaluated in such domains. For these reasons, in this paper we are motivated to evaluate how good character n-grams are in controversial opinions while considering the good precedent performance in opinion spam detection and the great difference between opinion spam and controversial opinions. Our purpose is to observe the suitability of character n-gram features for deception detection in domains very different from opinion spam, rather than to overcome all the baselines in controversial opinions.

2 Text Representation

Character n-grams are sequences of n characters present in the text, which are able to capture lexical and stylistic information. In this work, we evaluate different values for n and ten categories of n-grams with the purpose of finding

the most effective ones. These categories are related to three linguistic aspects: morpho-syntax, thematic content and style [10], as illustrated below:

- PREFIX: The first n characters of a word: "is killing".
- SUFFIX: The last n characters of a word: "is killing".
- SPACE_PREFIX: The first $n-1$ characters of a word, beginning with a space : "is killing".
- SPACE_SUFFIX: The last $n-1$ characters of a word ending with a space: "is killing".
- WHOLE_WORD: A whole word with n characters: "not moral".
- MID_WORD: n consecutive characters of a word, without the first and the last: "killing".
- MULTI_WORD: Include a space in the middle of the n-gram: "this person".
- BEG_PUNCT: A character n-gram where the first character is a punctuation: "essay. it would".
- MID_PUNCT: A character n-gram containing a punctuation mark: "essay. it".
- END_PUNCT: A character n-gram where only the last character is a punctuation: "essay.".

3 Experiments

3.1 Datasets Description

For both Abortion and Death Penalty domains, opiners were asked to express both the true opinion and the opposite on the topic, imaging that they were taking part in a debate. In the Best Friend domain, opiners were asked to write about their best friend and describe the detailed reasons for their friendship. Subsequently, they were asked to think about a person they could not tolerate, and describe her/him as if s/he were their best friend [9]. Table 1 shows the statistics of the three used datasets.

Table 1. Statistics for the datasets. Each domain has the information related to the deceptive (D) and truthful (T) classes: number of instances, the instances' vocabulary size, as well as the instances' length in characters and words.

	Instance		Length(ch)		Vocabulary		Length	
Domain	T	D	T	D	T	D	T	D
Abortion	100	100	499	359	64	50	101	73
Best friend	100	100	337	266	51	40	72	57
Death penalty	100	100	463	395	60	54	93	78

3.2 Experimental Setup

Preprocessing: We maintain all the characters present in the texts (e.g., punctuation marks, numbers, delimiters, etc.). The only normalization process was to convert all words to lowercase letter.
Feature Extraction and Selection: We considered char n-grams with n from 3 to 7, and discarded all features with a corpus frequency less than 3.
Classification: We used the Naïve Bayes (NB) and Support Vector Machine (SVM)[1] algorithms with a binary[2] weighting scheme.
Evaluation: We applied a 10-fold cross-validation procedure and used the accuracy as evaluation measure.

[1] SVM from *sklearn* with linear kernel, and default parameters.
[2] *tf* and *tf-idf* weighting also were used, but with binary weighting the classifiers achieve better results.

3.3 Results

This section presents the results achieved with character n-grams. We considered n-grams from 3 to 7, but character 5-grams showed slightly better performance taking into account the three datasets. Therefore, the following analyses were carried out with character 5-grams.

Table 2 shows the results obtained with character 5-grams, as well as the results from main related works. Interestingly, this simple representation is able to capture relevant information for detecting deception in controversial opinions. The results achieved with these simple sequences of characters are better than those obtained with a more complex, linguistically-motivated, representation using LIWC [9], which may be due to the fact that character n-grams combine information about style and content more specific to the domain at hand. However, approaches using *deep syntax* and LDA topics better discriminate between deceptive and truthful classes.

Table 2. Comparison of our results with other works on the same corpora. The classifier used by each author is given in the same cell as the accuracy.

Work	Abortion		Best friend		Death penalty	
words [7]	70%	NB	77%	SVM	67.4%	NB
LIWC [9]	73.03%	SVM	75.98%	SVM	60.36%	SVM
Deep syntax + words [4]	77%	SVM	85%	SVM	71%	SVM
LDA+words[6]	87.5%	SVN	87%	SVN	80%	SVN
character 5-grams	74%	NB	80.15%	SVM	63.95%	SVM

Qualitative Analysis. One single character n-gram can capture different things, for example, the 5-gram *count* can represent a prefix in *count*ry and a suffix in ac*count*. With the purpose of analyzing if char n-grams lose information for this phenomena, we divided them into the ten categories described in Sect. 2. Table 3 shows the three categories with the best results in at least one domain. In Abortion all the categories are almost equally important, while in Best Friend and Death Penalty the content and the way in which punctuation marks are used are the most important respectively.

Table 3. Accuracy with each category of character 5-grams using Naïve Bayes. Boldface indicates the highest value for each column.

Type of character 5-gram	Abortion	Best friend	Death Penalty
SPACE_SUFFIX	60%	**79%**	53%
BEG_PUNCT	60%	67%	**62%**
MULTI_WORD	66%	**79%**	60%
character 5-gram	**74%**	**79%**	54%

Table 4. Top 10 highest relevant features in the deceptive class. Each character 5-gram is highlighted in yellow and inserted in a context of each dataset.

(a) Relevant words

Abortion	Best Friend	Death Penalty
god	does	having
killing	this	him
babies	person	man
murder	his	practice
necessary	guy	rid
chance	nice	lesson
morally	how	around
evil	never	they
mistake	trustworthy	chance
innocent	wonderful	her

(b) Relevant character 5-grams

Abortion	Best Friend	Death Penalty
is murder	this person	convicted of
is murder	would never	killing another
deserves a	this person	man
is morally	this individual	having
is killing	this person	having
babies	he is	no matter
killing	he does	easy. it would
way of	this guy	no matter
babies to be	is a great	matter what
not moral	of his	need to

Another qualitative analysis was carried out to find the most relevant features. This was done evaluating the *mutual information* between features and classes (i.e. deceptive, truthful). Table 4 shows the 10 most representative words and character 5-grams used by deceivers in each topic. We observe that two differences among these representations arise from this table: (i) while one word could be taken as one feature in the word representation, many features are derived from the same word in character 5-grams, which is better for dealing with misspellings; (ii) the same character 5-gram can come from two different but related words, making it possible for two words that are semantically related to be reduced to a single feature, such is the case of *moral* in a*moral* and *moral*ly.

From the 5-grams given in Table 4b, and their respective contexts, we can draw the following conclusions: (i) deceivers tend to associate abortion with *murder* or *killing*, (ii) tend to distance themselves from their "best friend" (*this person/guy/individual*), and (iii) affirm their fake beliefs denying the importance of other factors. Additionally, we also noticed that in the Best Friend domain, deceivers tend to use expressions with *he is/does*, while non-deceivers use plural first person pronouns to talk about activities they do together; in truthful opinions is more common expressions that emphasize that what they say is what they really believe (e.g. *I believe that abortion is...; My honest opinion about...*); finally, non-deceivers tend to offer more detailed opinions.

4 Conclusions and Future Work

In this paper we addressed the problem of deceptive detection in controversial opinions using character n-grams as features. These features have not been studied in controversial opinions, although their simplicity and good results in opinion spam detection. Our experiments reported encouraging accuracies, between 63.95% and 80.15%, which suggest that character n-grams are effective for detecting deception in controversial domains, even better than using more complex representations based on linguistic features from LIWC. Character n-grams were able to capture shallow stylistic and thematic patterns not only useful

for the classification, but also for helping humans to analyze deceptive behaviors. According to their simplicity and performance, character n-grams are almost as satisfactory as more sophisticated representations. However, it seems that within our best results reported with character n-grams in these controversial domains, deep syntax and topic modeling must be considered in order to achieve high levels of accuracy.

In the future, we plan to analyze the relevant features obtained in all the domains and use them in cross-domain scenarios.

Acknowledgments. We would like to thank CONACyT for partially supporting this work under grants 613411, CB-2015-01-257383, and FC-2016/2410. The work of the last author was partially funded by the Spanish MINECO under the research project SomEMBED (TIN2015-71147-C2-1-P).

References

1. Aritsugi, M., et al.: Combining word and character n-grams for detecting deceptive opinions, vol. 1, pp. 828–833. IEEE (2017)
2. Buller, D.B., Burgoon, J.K.: Interpersonal deception theory. Commun. Theory **6**(3), 203–242 (1996)
3. Cagnina, L.C., Rosso, P.: Detecting deceptive opinions: intra and cross-domain classification using an efficient representation. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **25**(Suppl. 2), 151–174 (2017)
4. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection, pp. 171–175. Association for Computational Linguistics (2012)
5. Fusilier, D.H., Montes-y-Gómez, M., Rosso, P., Cabrera, R.G.: Detection of opinion spam with character n-grams. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9042, pp. 285–294. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18117-2_21
6. Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., Flores, J.J.G.: Cross-domain deception detection using support vector networks. Soft Comput. **21**(3), 1–11 (2016)
7. Mihalcea, R., Strapparava, C.: The lie detector: explorations in the automatic recognition of deceptive language. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 309–312. Association for Computational Linguistics (2009)
8. 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-Volume 1, pp. 309–319. Association for Computational Linguistics (2011)
9. Pérez-Rosas, V., Mihalcea, R.: Cross-cultural deception detection. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 440–445 (2014)
10. Sapkota, U., Solorio, T., Montes-y-Gómez, M., Bethard, S.: Not all character n-grams are created equal: a study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93–102 (2015)
11. Vrij, A.: Detecting Lies and Deceit: Pitfalls and Opportunities. Wiley, Hoboken (2008)