2017 IEEE 41st Annual Computer Software and Applications Conference

# Combining Word and Character N-grams for Detecting Deceptive Opinions

Al Hafiz Akbar Maulana Siagian
Computer Science and Electrical Engineering
Graduate School of Science and Technology
Kumamoto University
2-39-1 Kurokami, Chuo-ku, Kumamoto 860-8555, Japan
Center for Scientific Documentation and Information
Indonesian Institute of Sciences
Jl. Jend. Gatot Subroto 10, Jakarta 12710, Indonesia
Email: alha002@dbms.cs.kumamoto-u.ac.jp

Masayoshi Aritsugi
Big Data Science and Technology
Faculty of Advanced Science and Technology
Kumamoto University
2-39-1 Kurokami, Chuo-ku, Kumamoto 860-8555, Japan
Email: aritsugi@cs.kumamoto-u.ac.jp

***Abstract*—Essentially, opinion reviews are a valuable and trustworthy source of information for the readers. However, regarding the business purposes, a huge number of deceptive opinions are intentionally posted on the Web. In order to keep opinion reviews as a precious and trusted resource, we propose a method which focuses on detecting positive and negative deceptive opinions. In this paper, we explore the feasibility of combining word and character n-grams as a feature for detecting positive and negative deceptive opinions. The majority of studies in this task show that employing n-grams, i.e., words or characters, as a feature is sufficient to obtain good results. We examine our proposed method using a corpus about hotel reviews containing positive and negative opinions both deceptive and truthful. We extract each of word and character n-grams from reviews in the dataset, and then combine them as a feature. Our experiment results show that our proposed method outperforms methods of using the word or character n-grams alone. Furthermore, we consider applying the Principal Component Analysis (PCA) to classify the dominant and nonessential feature attributes for decreasing the size of feature attributes. The obtained results by removing the irrelevant feature attributes are similar to those of using all feature attributes.**

***Index Terms*—deceptive opinions; opinion spam; n-grams; feature reduction; fake reviews; combining word and character;**

## I. INTRODUCTION

Nowadays, a huge number of product or service providers run their business by online on the Web. Commonly, people who have ever purchased by online will give a review as a feedback not only for the companies but also for next potential customers. Therefore, many customers utilize online reviews as their pre-purchase consideration. Rudolph [1] reported that 90% of the consumers look and rely on online reviews before making the decision to use products or services.

Unfortunately, regarding the business purposes, this phenomenon is abused by generating deceptive opinions to mislead the consumers [2]. Since the deceptive opinions themselves seem as well as the truthful opinions, it is very difficult to distinguish between them. Particularly, the deceptive opinions are divided into two categories, namely, positive and negative deceptive opinions. Positive deceptive opinions are reviews to upgrade a low-quality product or service, while negative deceptive opinions are an action to degrade the excellent ones [3]. Thus, detecting deceptive opinions is very complicated and challenging issues.

First attempt to detect the deceptive opinions, or called opinion spams, was done by [4]. They inferred that detecting the deceptive opinions is very difficult and nearly impossible to identify manually. Mukherjee et al. [5] explained that it is almost improbable to assign the authenticity of the reviews, especially in the real case. [6] also confirms that employing a human as a judge is difficult to distinguish between truthful and deceptive opinions.

Furthermore, Ott et al. [7], [8] conducted intensive studies by using the human and automatic system to classify the untruthful reviews. They concluded that utilizing the automated classifier obtained better results than the human judged. They also released a gold-standard dataset which contains positive and negative deceptive opinion instances. By using the gold-standard dataset, it was possible to perform a supervised method and trustworthy evaluation for this difficult task as shown in [9]. Interestingly, by employing only n-grams features, i.e., word and character n-grams, it was able to achieve promising results.

According to the encouraging results of [9] using n-grams features only, we have been motivated to explore another possibility of employing the text features, that is, combinations of word and character n-grams. In this paper, we evaluate our proposed method by employing the gold-standard dataset[1] invented by [7], [8]. The first experiment results show that the combination of word and character n-grams as a feature outperformed the results of without combination.

On the other hand, there is one primary issue when dealing with text classification tasks, that is, the huge number of feature attributes. Consequently, we need high computational costs to deal with it. Moreover, the machine learning algorithm becomes more difficult to analyze and classify the reviews.

[1] http://www.myleott.com/op_spam/

0730-3157/17 $31.00 © 2017 IEEE
DOI 10.1109/COMPSAC.2017.90

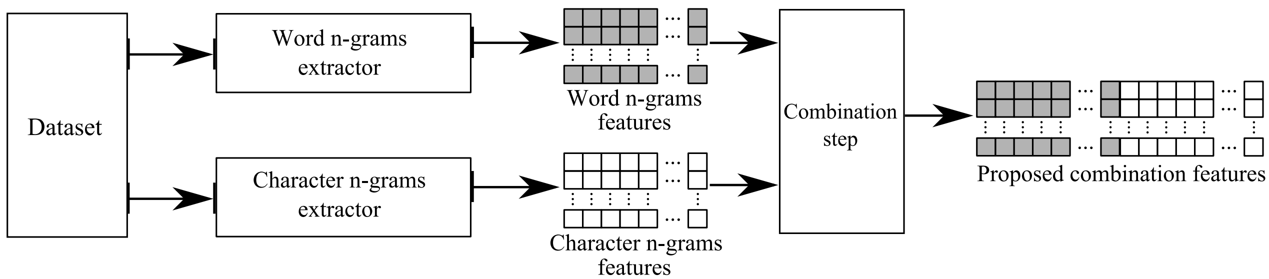



Fig. 1. Combination of word and character n-grams features process.

Lim et al. [10] tried to avoid those problems by proposing rating behavior features to detect product review spammers.

Decreasing of the feature set by removing non-essential feature attributes is a fundamental part of opinion classification tasks [11], [12], [13]. Cagnina and Rosso [14] tried to reduce the size of the feature attributes in this task by employing the Linguistic Inquiry and Word Count (LIWC). To address that problem, we consider applying the Principal Component Analysis (PCA) as our feature attributes reduction technique in this work. The last experiment results show by removing nonessential feature attributes we were able to achieve similar results to the original results, even better.

Concretely, this paper makes mainly the following contributions:

- We propose a method to combine word and character n-grams as a feature for detecting positive and negative deceptive opinions.
- We consider applying the Principal Component Analysis (PCA) as a feature attributes reduction technique in this study.
- Our experimental results show that the combination of word and character n-grams as a feature and PCA could perform well in detecting deceptive opinions.

The remainder of this paper is organized as follows. Section II presents some related work. Section III describes our method. We explain the dataset and experimental setup in Section IV. Section V exhibits the experiments for examining our approach and shows some results. Finally, Section VI provides the conclusion and direction for future work.

## II. Related Work

Regarding the difficulties of conducting a study in the real case, many studies in opinion spam detection rely on the synthetic dataset [15]. Ott et al. [7], [8] generated the gold-standard dataset of deceptive and truthful for hotel reviews. Another study [16] tried to produce the new gold-standard dataset which covers for three domains, i.e., hotel, restaurant and doctor reviews. However, they mainly used word n-grams, i.e., unigrams, bigrams, and trigrams, as their features. Fusilier et al. [3], [17] employed word n-grams as features to evaluate their novel PU-learning approach for detecting positive and negative deceptive opinions. Ott et al. [7], [8] and Li et al. [16] performed word n-grams by combining other approaches such as LIWC and Part of Speech (POS)as their features.

There are two studies [18], [19] employing character n-grams as a feature. Nevertheless, they focused on the authorship attribution. On the other hand, Blamey et al. [20] and Kanaris et al. [21] tried comparing between word and character n-grams as features. However, their objectives were to classify the sentiment analysis and identify the email spam.

Based on the study by [11], PCA is an effective method to reduce the number of feature attributes in the opinion mining task. The PCA was able to select the irrelevant and dominant feature attributes. Still, their concentration was on the sentiment analysis classification in the opinion reviews, i.e., not deceptive opinions. Further study about the PCA in the text classification task has been done by [13]. They employed PCA as a based method to propose a novel feature reducing technique. Nevertheless, their study focused on email classifications.

## III. Method

In this section, we propose a combination of word and character n-grams as a feature. We make use of the PCA as our feature attributes reduction method.

### A. Combination of Word and Character N-grams as a Feature

For creating each word and character n-grams feature, we extracted them by using the corpus by Ott et al. [7], [8]. Particularly, we follow a bag of word n-grams and a bag of character n-grams approaches by selecting n contiguous words or characters from a given sentence of reviews.

First, we create word n-grams features, i.e., unigrams, bigrams, and trigrams, from reviews in the dataset. Then, we construct character n-grams features by considering 1 to 5 grams only, i.e., including the space character, from the dataset. Our motivation is based on the study by [9]. They observed 1 to 500 n-grams and their best results were achieved by using 5 and 4 character n-grams features for positive and negative opinion instances, respectively.

As illustrated in Fig. 1, we examine the feasibility of combining the word and character n-grams with respect to the results reported in [9]. Since our proposal is to examine the feasibility of combining word and character n-grams as a feature, we incorporate them without any other considerations.

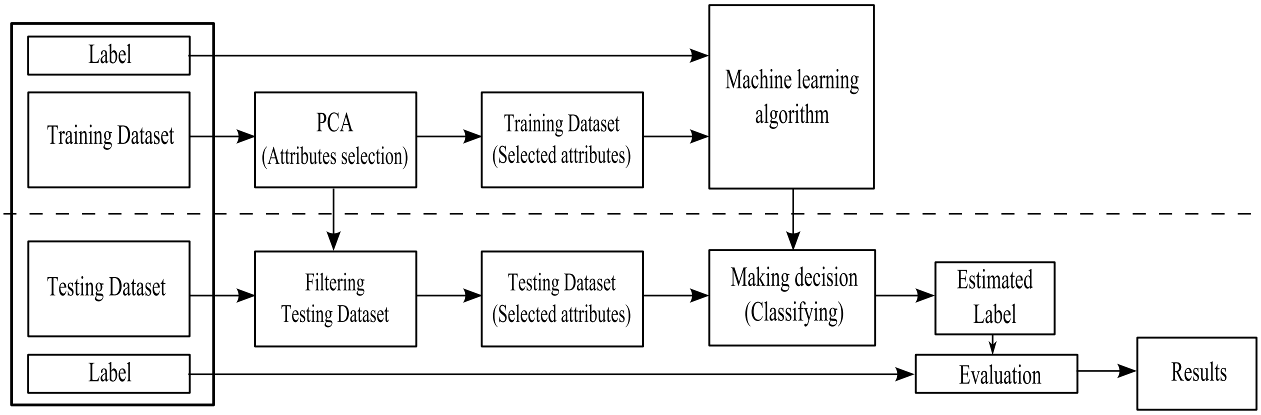



Fig. 2. General overview of our method.

In other words, we combine each one of the word n-grams with each one of the character n-grams to generate a new feature at the combination step. Then, we provide the new feature to create the training and testing dataset for the classifier.

We assume that using the combination of word and character n-grams as a feature would provide more precise features and cover the lack of each feature for the machine learning process. We expect that our proposed method would tackle the unusual text pattern problem, e.g., typo, in the real case.

### B. Feature Attributes Reduction

Datasets used for classification processing such as text and image classifications are, mostly, of enormous dimensions. Discovering another smaller dimension that still represents the original one is an important part of learning algorithms [11], [22], [23]. In particular, this process is for simplifying the classification step.

Typically, the size of character n-grams feature attributes is smaller than the size of word n-grams feature attributes. However, because we combine them to generate a new feature, naturally, the size of our proposed feature attributes is slightly larger than the size of individual of character or word n-grams feature attributes.

To solve those issues in this study, we utilize the PCA for reducing the number of feature attributes. Specifically, it is able to remove non-essential features but still reflects the genuine ones. We illustrate the PCA algorithm as follows:

1) Assume A is an (n x m) matrix, where n = #reviews, and m = #feature attributes.
2) Calculate the covariance matrix.
3) Calculate eigenvalues and eigenvectors.
4) Calculate the standardized transformation matrix.
5) Calculate the weight of feature attributes for each review. The range is from 0 to 1.
6) Select feature attributes.

### C. Deceptive Opinion Detection

We exploit the machine learning technique with employing features extracted by the method described in Section III-A. We provide the training data to be learned by machine learning for labeling the testing data. Then, we did the evaluation process by using the estimated label from the classifying step. We illustrate the general overview of our method in Fig. 2.

## IV. Dataset and Experimental Setup

### A. Dataset

To evaluate our proposed method, we use the gold-standard dataset provided by Ott et al. [7], [8] about hotel reviews located in Chicago. The dataset contains 1600 labeled instances of positive and negative opinions for both deceptive and truthful reviews.

### B. Experimental Setup

We performed all the experiments by employing Octave[2] application. We also did some experimental settings to support and facilitate our experiments as follows:

1) Tokenization preparation: We followed the dataset preprocessing method conducted by [3] and [9] that used alphabetical tokens only. Particularly, we expunged the punctuation and numerical sign. Then, we altered all letters into the lowercase.
2) Classification algorithms: We utilized the nan package[3] in Octave for the machine learning algorithm. We performed the experiments using several classification algorithms, e.g., Regression Analysis (REG), Support Vector Machines (SVM), SVMlib, PSVM, NBC, and aNBC. However, only REG consistently showed the best results among them. It is important to mention that the training time for the machine learning process was approximately under 5 minutes for each approach.
3) Experimental measurement: We measured our proposed method by calculating the precision, recall, and F-measure of each deceptive and truthful review. To exhibit our experimental results, we computed and used

[2]https://www.gnu.org/software/octave/
[3]https://octave.sourceforge.io/nan/index.html

the macro average of F-measure for both positive and negative opinions. The F-measure, recall, and precision for each opinion category are computed as follows:

$$\text{precision} = \frac{tp}{tp + fp} = \frac{tn}{tn + fn} \tag{1}$$

$$\text{recall} = \frac{tp}{tp + fn} = \frac{tn}{tn + fp} \tag{2}$$

$$\text{F-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

4) The distribution for training and testing dataset: We used 640 reviews and 160 reviews both deceptive and truthful as the training and testing dataset respectively for each of positive and negative opinions. These distributions correspond to the 5 fold cross validation technique which we used in this study.

## V. Experiments

### A. Experiment 1: Combination of Word and Character N-grams as a Feature

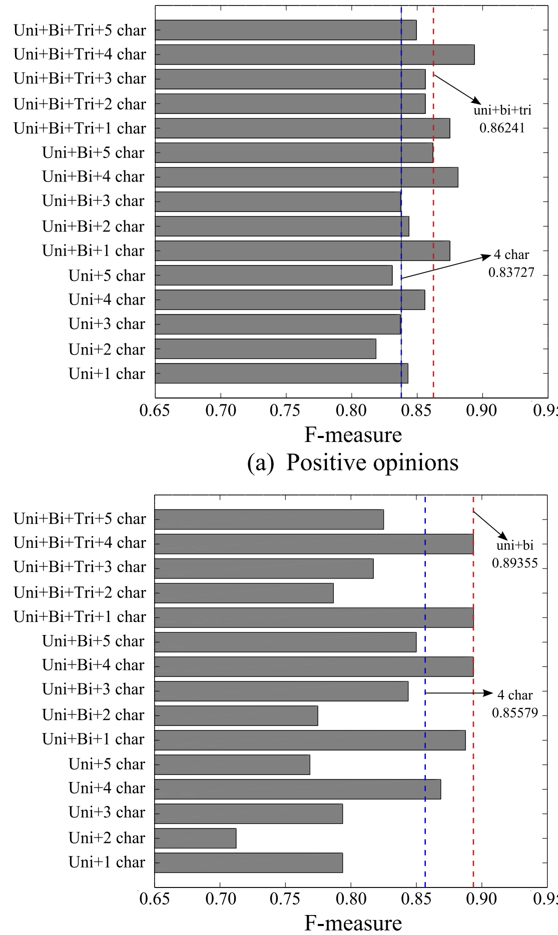
The primary goal of this experiment was to analyze the feasibility of using the combination of word and character n-gram as a feature. Results in Fig. 3 clearly show that using the combination of word and character n-grams as a feature was consistently able to achieve better results than employing the word or character n-grams as features only. Specifically, the unigrams + bigrams + trigrams + 4 characters feature following by a combination of unigrams + bigrams + trigrams + 1 character as features generated the best results for positive and negative opinions respectively.

We present the amount of feature attribute sizes and values of macro average F-measure for each combination of word and character n-grams features on both polarities in Table I. It is important to note that results by employing a combination of unigrams + bigrams + trigrams + 1 character as features were able to exceed the best results of each word and character n-grams feature. Surprisingly, it can reach the highest point for the negative opinion instance.

Another interesting observation from Table I is the obtained results using a combination of word and characters as features exhibit increasing trends. Although results of the character n-grams features alone showed a low significant discrepancy compared to the word n-grams ones, it contributed to give better outcomes as shown in Table I. Particularly, it also confirms that employing only n-grams features is sufficient to obtain good results.

### B. Experiment 2: The Effectiveness of Combining Word and Character N-grams

Generally, there could be mistyping in product reviews on the Web. In order to resemble real cases, we examined the effectiveness of our proposed method by making a modification in text reviews of the testing data. We expected by altering some words or compositions in the testing dataset, i.e., not available in the database of features, it would be more



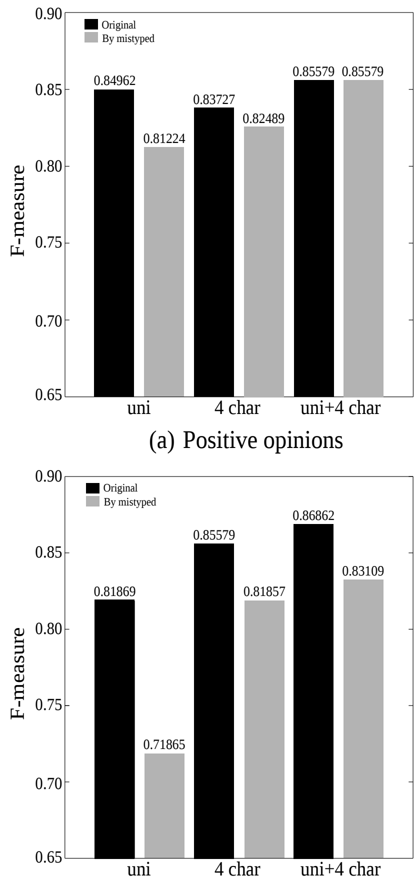

(a) Positive opinions

(b) Negative opinions

Fig. 3. Results obtained using our proposed features. The red dash line stands for the best result of the baseline of word n-grams, and the blue dash line represents the best result of the baseline of character n-grams.

TABLE I
Results by using the combination of word and character n-grams as features in positive and negative opinions. Results in bold indicate the best performance for each polarity.

| FEATURES | POSITIVE | | NEGATIVE | |
|---|---|---|---|---|
| | size | F-measure | size | F-measure |
| uni+1 char | 5810 | 0.84301 | 7879 | 0.79368 |
| uni+2 char | 6307 | 0.81857 | 8389 | 0.71232 |
| uni+3 char | 10075 | 0.83709 | 12447 | 0.79368 |
| uni+4 char | 22496 | 0.85579 | 26936 | 0.86862 |
| uni+5 char | 49615 | 0.83093 | 61687 | 0.76874 |
| uni+bi+1 char | 43442 | 0.87498 | 63765 | 0.88750 |
| uni+bi+2 char | 43939 | 0.84370 | 64275 | 0.77468 |
| uni+bi+3 char | 47707 | 0.83727 | 68333 | 0.84374 |
| uni+bi+4 char | 60128 | 0.88121 | 82822 | 0.89355 |
| uni+bi+5 char | 87247 | 0.86196 | 117573 | 0.84991 |
| uni+bi+tri+1 char | 113680 | 0.87492 | 170592 | **0.89375** |
| uni+bi+tri+2 char | 114177 | 0.85620 | 171102 | 0.78667 |
| uni+bi+tri+3 char | 117945 | 0.85620 | 175160 | 0.81714 |
| uni+bi+tri+4 char | 130366 | **0.89375** | 189649 | 0.89355 |
| uni+bi+tri+5 char | 157485 | 0.84941 | 224400 | 0.82500 |




(a) Positive opinions

(b) Negative opinions

Fig. 4. Results of unigrams, 4 characters, and the combination of unigrams + 4 characters features by comparing the original results to the results of after applying mistypes on the testing dataset.

difficult for the machine learning algorithm to classify the deceptive opinions.

For simplicity, we randomly manipulated some unigrams feature attributes, i.e., changing original words into random text and eliminating spaces, in reviews of the testing dataset. We considered this test on unigrams, 4 characters, and unigrams + 4 characters n-grams features.

Our experimental results in Fig. 4 show that the combination of unigrams + 4 characters n-grams features consistently outperformed the others for both positive and negative instances. Although each of unigrams and 4 characters results decreased, it did not give a meaningful impact on influencing our proposed method results.

According to this experiment result, we analyze that the combination of word and character n-grams as a feature provides more comprehensive features for the machine learning process. Combining the word and character n-grams as a feature is able to cover the unusual content pattern in text reviews, e.g., mistype. We noticed that this new feature would be possible to classify better positive and negative deceptive opinions. Nevertheless, it would be better to carry out further detail analysis about the causes of this second experimental results.

### *C. Experiment 3: Applying the PCA on Selected N-grams Features*

The principal purpose of this experiment is to inspect the feasibility to reduce the number of irrelevant feature attributes for simplifying the machine learning task. For executing this analysis, we used the datasets and features as well as in experiment 1. Particularly, we just added the feature attributes selection process by employing the PCA after preparing the training data. However, for better understanding, we just showed for several selected features which achieved optimal results in experiment 1.

Tables II and III present the results obtained by utilizing the PCA technique for positive and negative opinions respectively.

TABLE II
RESULTS BY APPLYING THE PCA ON SELECTED N-GRAMS FEATURES IN POSITIVE OPINIONS. BOLD SIGNS IN SELECTION COLUMNS INDICATE SIMILAR OR BETTER RESULTS COMPARED TO THE ORIGINAL RESULTS FOR EACH N-GRAMS FEATURE.

| Positive opinions | | | | |
|---|---|---|---|---|
| FEATURES | | Original | Selection | |
| uni+bi | size | 43415 | 36605 | 10482 |
| | F-measure | 0.85611 | **0.85611** | 0.85597 |
| uni+bi+tri | size | 113653 | 93093 | 16834 |
| | F-measure | 0.86241 | **0.86241** | 0.86231 |
| 4 char | size | 16713 | 15645 | 11420 |
| | F-measure | 0.83727 | **0.83727** | **0.84345** |
| uni+bi+1 char | size | 43442 | 36632 | 10509 |
| | F-measure | 0.87498 | **0.87498** | 0.86241 |
| uni+bi+4 char | size | 60128 | 52250 | 21902 |
| | F-measure | 0.88121 | **0.88121** | 0.85597 |
| uni+bi+tri+1 char | size | 113680 | 93120 | 16861 |
| | F-measure | 0.87492 | **0.87492** | 0.86241 |
| uni+bi+tri+4 char | size | 130366 | 108738 | 28254 |
| | F-measure | 0.89375 | **0.89375** | 0.86862 |

TABLE III
RESULTS BY APPLYING THE PCA ON SELECTED N-GRAMS FEATURES IN NEGATIVE OPINIONS. BOLD SIGNS IN SELECTION COLUMNS INDICATE SIMILAR OR BETTER RESULTS COMPARED TO THE ORIGINAL RESULTS FOR EACH N-GRAMS FEATURE.

| Negative opinions | | | | |
|---|---|---|---|---|
| FEATURES | | Original | Selection | |
| uni+bi | size | 63738 | 53528 | 15116 |
| | F-measure | 0.89355 | **0.89355** | 0.86870 |
| uni+bi+tri | size | 170565 | 139068 | 24448 |
| | F-measure | 0.88722 | **0.88722** | 0.88121 |
| 4 char | size | 19084 | 18016 | 13311 |
| | F-measure | 0.85579 | **0.85579** | **0.85579** |
| uni+bi+1 char | size | 63765 | 53555 | 15143 |
| | F-measure | 0.88750 | **0.88750** | 0.84374 |
| uni+bi+4 char | size | 82822 | 71544 | 28427 |
| | F-measure | 0.89355 | **0.89355** | **0.89355** |
| uni+bi+tri+1 char | size | 170592 | 139095 | 24475 |
| | F-measure | 0.89375 | **0.89375** | 0.86250 |
| uni+bi+tri+4 char | size | 189649 | 157084 | 37759 |
| | F-measure | 0.89355 | **0.89355** | 0.83750 |

Although we used less number of feature attributes, results in the selection columns were similar, sometimes better, compared to the original results. These matters were not only in positive opinions but also in negative opinions. Based on the results of this experiment, we observed that removing the irrelevant feature attributes, i.e., zero weight, has no significant impact. Therefore, we can conclude that employing the PCA would be effective and precious as a feature attribute reduction technique.

## VI. Conclusion

In this study, we explored the feasibility of combining word and character n-grams as features to detect positive and negative deceptive opinions. We employed the gold-standard dataset provided by Ott et al. [7], [8], which contains deceptive opinions both positive and negative classes, to evaluate our proposed method. The experimental results showed that the combination of word and character n-grams as features consistently outperformed the word and the character n-grams features. Furthermore, we investigated the effectiveness of our proposed method to deal with real case problem of text in reviews, e.g., typo. The obtained results exhibited that the combination of word and character n-grams as a feature was able to overcome the unusual text pattern. As an additional contribution, we examined the PCA as a feature attributes reduction technique. The results obtained by removing the irrelevant feature attributes were similar to, or even better than, the results of using all feature attributes.

For future work, we intend to utilize other classification algorithms and feature attributes reduction techniques such as Random Forest and kernel PCA. It also would be worth to take into account the computational time of our work. Furthermore, we aim to use larger datasets, such as in [16], so that we could examine our proposed method more comprehensively, e.g., by performing 5 and 10 fold cross validation methods.

## References


[1] S. Rudolph, "The impact of online reviews on customers' buying decisions [Infographic]," Business 2 Community, 2015. [Online]. Available: http://www.business2community.com/infographics/impact-online-reviews-customers-buying-decisions-infographic-01280945. [Accessed: Jan. 9, 2017].

[2] A. Topping, "Historian Orlando Figes agrees to pay damages for fake reviews," in The Guardian, 2010. [Online]. Available: https://www.theguardian.com/books/2010/jul/16/orlando-figes-fake-amazon-reviews. [Accessed: Jan. 9, 2017].

[3] D. H. Fusilier, M. Montes-y-Gómez, P. Rosso, and R. G. Cabrera, "Detecting positive and negative deceptive opinions using PU-learning," Information Processing & Management, vol. 51, no. 4, pp. 433-443, Jul. 2015.

[4] N. Jindal and B. Liu, "Opinion spam and analysis," in Proceedings of the international conference on web search and data mining - WSDM '08, pp. 219-230, 2008.

[5] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, "Detecting group review spam," in Proceedings of the 20th international conference companion on world wide web - WWW '11, pp. 93-94 2011.

[6] C. Harris, "Detecting deceptive opinion spam using human computation," in Proceedings of the 4th AAAI Human Computation Workshop - HCOMP '12, pp. 87-93, 2012.

[7] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock, "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. pp. 309-319, 2011.

[8] M. Ott, C. Cardie, and J.T. Hancock, "Negative deceptive opinion spam," in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 497-501, 2013.

[9] D. H. Fusilier, M. Montes-y-Gómez, P. Rosso, and R. G. Cabrera, "Detection of opinion spam with character n-grams," in Computational Linguistics and Intelligent Text Processing - CICLing '16, Springer, pp. 285-294, 2015.

[10] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proceedings of the 19th ACM international conference on information and knowledge management - CIKM '10, pp. 939-948, 2010.

[11] G. Vinodhini and R. M. Chandrasekaran, "Opinion mining using principal component analysis based ensemble model for e-commerce application," CSI Transactions on ICT, vol. 2, no. 3, pp. 169-179, Sep. 2014.

[12] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages," ACM Transactions on Information Systems, vol. 26, no. 3, pp. 1-34, Jun. 2008.

[13] J. C. Gomez and M. F. Moens, "PCA document reconstruction for email classification," Computational Statistics & Data Analysis, vol. 56, no. 3, pp. 741-751, Mar. 2012.

[14] L. Cagnina and P. Rosso, "Classification of deceptive opinions using a low dimensionality representation," in Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 58-66, 2015.

[15] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," Journal of Big Data, vol. 2, no. 1, pp. 2-23, Oct. 2015.

[16] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1566-1576, Jun. 2014.

[17] D. H. Fusilier, R. G. Cabrera, M. Montes-y-Gómez, and P. Rosso, "Using PU-learning to detect deceptive opinion spam," in Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 38-45. 2013.

[18] U. Sapkota, S. Bethard, M. Montes-y-Gómez, and T. Solorio, "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.

[19] E. Stamatatos, "On the robustness of authorship attribution based on character n-gram features," Journal of Law and Policy, vol. 21, no. 2, pp. 421-439, 2013.

[20] B. Blamey, T. Crick, and G. Oatley, "R U :-) or :-( ? character- vs. word-gram feature selection for sentiment classification of OSN corpora," in Proceedings of the 32nd SGAI international conference on innovative techniques and applications of artificial intelligence, Springer, pp. 207-212, 2012.

[21] I. Kanaris, K. Kanaris, I. Houvardas, and E. Stamatatos, "Word versus character n-grams for anti-spam filtering," International Journal on Artificial Intelligence Tools, vol. 16, no. 06, pp. 1047-1067, Dec. 2007.

[22] E. Liberty, "Simple and deterministic matrix sketching," in Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '13, pp. 581-588, 2013.

[23] M. Liu, K. Liu, and X. Ye, "Find the intrinsic space for multiclass classification," in Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies - ISABEL '11, ACM, pp. 106:1-106:5, 2011.