Linguistic Features to Identify Extreme Opinions: An Empirical Study

Sattam Almatarneh(✉) and Pablo Gamallo

Centro Singular de Investigación en Tecnoloxías da Información (CITIUS),
Universidad de Santiago de Compostela, Rua de Jenaro de la Fuente Domínguez,
15782 Santiago de Compostela, Spain
{sattam.almatarneh,pablo.gamallo}@usc.es

Abstract. Studies in sentiment analysis and opinion mining have examined how different features are effective in polarity classification by making use of positive, negative or neutral values. However, the identification of extreme opinions (most negative and most positive opinions) have overlooked in spite of their wide significance in many applications. In our study, we will combine empirical features (e.g. bag of words, word embeddings, polarity lexicons, and set of textual features) so as to identify extreme opinions and provide a comprehensive analysis of the relative importance of each set of features using hotel reviews.

Keywords: Sentiment analysis · Opinion mining
Linguistic features · Classification · Extreme opinion

1 Introduction

The information revolution is the most prominent feature of this century. The world has become a small village especially with the proliferation of social networking sites where anyone in the world can sell, buy or express their opinion. The vast amount of information on the Internet has become a source of interest for studies, as it offers an excellent opportunity to extract information and organize it according to the need. In the last two decade, an immense number of studies have been carried in the field of opinion mining and sentiment analysis. The main task in Opinion Mining is polarity classification, which occurs when a piece of text stating an opinion is classified into a predefined set of polarity categories (e.g., positive, neutral, negative). Reviews such as “like” versus “dislike” are examples of two-class polarity classification. An unusual way of performing sentiment analysis is to detect and classify extreme opinions, which represent the most negative and most positive opinions about a topic, an object or an individual. An extreme opinion is the worst or the best view, judgment, or appraisal formed in ones mind about a particular matter.

One of the main motivations for detecting extreme opinions is the fact that they actually stand for *pure* positive and negative opinions. As rating systems have no clear borderlines on a continuum scale, weakly polarized opinions

H. Yin et al. (Eds.): IDEAL 2018, LNCS 11314, pp. 215–223, 2018.
https://doi.org/10.1007/978-3-030-03493-1_23

(e.g. those rated as 4 and 2 in a 1 to 5 rating system) may be in fact closer to neutral statements. According to Pang and Lee [11], “it is quite difficult to properly calibrate different authors’ scales, since the same number of *stars* even within what is ostensibly the same rating system can mean different things for different authors”. Given that rating systems are defined on a subjective scale, only extreme opinions can be seen as natural, transparent, and non ambiguous positive or negative statements. Extreme opinions only constitute a small portion of the opinions on Social Media. According to [11], only about 5% of all opinions are on the most extreme points of a scale, which makes the search for these opinions a challenge. We are then confronted with a challenging task.

It is not surprising that extreme views have a strong impact on product sales, since they influence customer decisions before buying. Previous studies analyzed this relationship, such as the experiments reported in [8], which found that as the high proportion of negative online consumer reviews increased, the consumer’s negative attitudes also increased. Another motivation for the identification of extreme opinions is the current use of bot technology by cyborgs on social networks. These bots are designed to sell products or attract clicks, amplifying false or biased stories in order to influence public opinion.

The main objective of this article is to examine the effectiveness and limitations of different linguistic features to identify extreme opinions in the hotels’ reviews. Our main contribution is to report an extensive set of experiments aimed to evaluate the relative effectiveness of different linguistic features for two binary classification tasks:

- very negative *vs.* not very negative opinions
- very positive *vs.* not very positive opinions

The rest of the paper is organized as follows. In the following Sect. 2, we describe the related work. Then, Sect. 3 describes the method. Experiments are introduced in Sect. 4, where we also describe the evaluation and discuss the results. We draw the conclusions and future work in Sect. 6.

2 Related Work

There are two main approaches to find the sentiment polarity at a document level. First, machine learning techniques based on training corpora annotated with polarity information and, second, strategies based on polarity lexicons. The success of both methods mainly depends on the choice and extraction of the proper set of features used to identify sentiments. There is a great number of surveys and books in sentiment analysis describing the main methods and comparing the usefulness of different linguistic and textual features. For instance, the most salient linguistic features for sentiment classification are listed in Chapter 3 of [9] book. [4] presented a systematic study of different sentence features for two tasks in sentiment classification: namely, polarity classification and subjectivity classification. [7] introduced a new approach to build fixed length vectors for paragraph, sentence, and document representation. [17] proposed an approach to find the polarity of reviews by converting text into numeric matrices using *countvectorizer* and TF-IDF, and then using them as input in machine learning algorithms for classification. Moreover, sentiment words are the core component in opinion mining and have been used in many studies. [10] built a lexicon containing a combination of sentiment polarity (positive, negative) with one of eight possible emotion classes (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) for each word. As far as we know, excerpted of our previous studies [2,3] no previous work has been focused on detecting extreme opinions. Our proposal, therefore, may be considered to be the first step in that direction.

3 Method

We deal with two document-level binary classification tasks: (1) very negative *vs.* not very negative, and (2) very positive *vs.* not very positive. These tasks can be achieved by automatic classifiers composed of training data in a supervised strategy. The characteristics of documents will be encoded as features in vector representation. These vectors and the corresponding labels feed the classifiers. In the experiments described later, we will examine the following sets of features:

- **N-grams Features:** We deal with n-grams based on the occurrence of unigrams and bigrams of words in the document. Unigrams (1g) and bigrams (2g) are valuable to detect specific domain-dependent (opinionated) expressions. The influence of this type of content features has been confirmed by several opinion mining studies [12,19]. We assign a weight to all terms by using two representations: Term Frequency-Inverse Document Frequency (TF-IDF) and CountVectorizer. TF-IDF is computed in Eq. 1.

$$tf/idf_{t,d} = (1 + log(tf_{t,d})) \times log(\frac{N}{df_t}). \tag{1}$$

  where $tf_{t,d}$ in the term frequency of the term t in the document d, N is the number of documents in the collection and, df_t is the number of documents in the collection containing t. CountVectorizer transforms the document to token count matrix. First, it tokenizes the document and according to a number of occurrences of each token, a sparse matrix is created. In order to create the Matrix, all stop words are removed from the document collection. Then, the vocabulary is cleaned up by eliminating those terms appearing in less than 4 documents to eliminate those terms that are too infrequent. To convert the reviews to a matrix of TF-IDF features and to a matrix of token occurrences, we used sklearn feature extraction python library.[1]
- **Doc2Vec:** We used the Doc2vec algorithm introduced in [7] to represent the reviews. This neural-based representation has been shown to be efficient when dealing with high-dimensional and sparse data [5,7]. Doc2vec learns features

[1] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

from the corpus in an unsupervised manner and provides a fixed-length feature vector as output. Then, the output is fed into a machine-learning classifier. We used a freely available implementation of the doc2vec algorithm included in gensim,[2] which is a free Python library. The implementation of the doc2vec algorithm requires the number of features to be returned (length of the vector). So, we performed a grid search over the fixed vector length 100.

- **Set of Textual Features (SOTF):** Many textual features may be used as evidences to detect extreme views: both very positive or very negative alike. In this study, we have extracted some of them to examine to what extent they influence the identification of extreme views. Uppercase characters may indicate that the writer is very upset or affected, so we counted the number of words written in uppercase letters. Also, intensifier words could be a reliable indicator of the existence of extreme views. So, we considered words such as mostly, hardly, almost, fairly, really, completely, definitely, absolutely, highly, awfully, extremely, amazingly, fully, and so on. Furthermore, we took into account negation words such as no, not, none, nobody, nothing, neither, nowhere, never, etc. In addition, we also considered elongated words and repeated punctuation such as (*sooooo, baaaaad, woooow, gooood, ???, !!!!,...etc*). These textual features have been shown to be effective in many studies related to polarity classification such as [6,16].
- **Sentiment Lexicons:** Sentiment words also called opinion words are considered the primary building block in sentiment analysis as it is an essential resource for most sentiment analysis algorithms, and the first indicator to express positive or negative opinions. In our previous studies, we described a strategy to build sentiment lexicons from corpora [1,3]. In this study, we used the same method to create two lexicons of the most negative words and another one for the most positive for hotels domain. VERY-NEG is a lexicon made up of words classified as MN or NMN, while VERY-POS is another lexicon consisting of words classified as MP or NMP[3]. The new sentiment lexicons for hotels were built from the text corpora introduced in [14,15]. The corpora[4] consist of online reviews collected from IMDB, Goodreads, OpenTable and Amazon/Tripadvisor. We only use the hotels and restaurants reviews from OpenTable an Tripadvisor. As shown in Table 1, we included lexicon-based features in the two classification tasks as follows. For MN *vs* NMN We represented the number of MN and the number of NMN terms in the document. We also included the proportion of MN and NMN terms. And the same way for the second classification task (MP *vs* NMP) We represented the number of MP and the number of NMP terms in the document. We also included the proportion of MP and NMN terms.

Table 1 summarizes all the features introduced above with a brief description for each one.

[2] https://radimrehurek.com/gensim/.
[3] https://github.com/citiususc/VERY-NEG-and-VERY-POS-Lexicons.
[4] http://www.stanford.edu/~cgpotts/data/wordnetscales/.

Table 1. Description of all the considered linguistic features in order to identify the most negative opinions (MN vs. NMN) and the most positive opinions (MP vs. NMP)

Features	Descriptions
N-grams	Unigram TF-IDF(1g) Unigram CountVectorizer(1g) Unigram and Bigram TF-IDF (1g 2g) Unigram and Bigram CountVectorizer (1g 2g)
Doc2Vec (100 feat.)	Generate vectors for the document
SOTF (8 feat.)	Number and proportion of negation words in the document Number and proportion of uppercase words in the document Number and proportion of elongated words and punctuations in the document Number and proportion of intensifiers words in the document
VERY-NEG (4 feat.)	Number and proportion of MN terms in the documents Number and proportion of NMN terms in the documents
VERY-POS (4 feat.)	Number and proportion of MP terms in the documents Number and proportion of NMP terms in the documents

4 Experiments

4.1 Data collection

In order to extract extreme opinions, we require to analyze document collections with scaled opinion levels (e.g. rating) and extract those documents associated with the lowest and highest scale. We obtained our dataset from Expedia crowd-sourced data. The HotelExpedia dataset[5] originally contains 6030 hotels and 381941 reviews from 11 different hotel locations. The datasets are cleaned and prepared for analysis by applying the following three preprocessing steps: (1) data deduplication operation is performed in order to remove such duplicate reviews; (2) 3-stars reviews were deleted since they tend to contain neutral views; (3) all reviews containing less than three words and blank reviews were also removed. After the above three data cleansing operations, the final datasets consists of 20,000 reviews, being 5,000 for each category: 1, 2, 4 and 5 stars.

[5] http://ave.dee.isep.ipp.pt/~1080560/ExpediaDataSet.7z.

4.2 Training and Test

Since we are facing a text classification problem, any existing supervised learning method can be applied. Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization [12]. We decided to utilize *scikit*[6] which is an open source machine learning library for Python programming language [13]. This library implements several classifiers, including regression and clustering algorithms. We chose SVMs as our classifier for all experiments, hence, in this study we will only summarize and discuss results for this learning model. The dataset was randomly partitioned into training (75 %) and test (25 %). In our analysis, we employed 5_fold cross_validation and the effort was put on optimizing F1 which is computed with respect to MN and MP (which is the target class). We also measured statistical significance with a paired, two-sided micro sign test [18]. This is a statistical method to test for consistent differences between pairs of observations based on their binary decisions on all the document/category pairs, and it applies the Binomial distribution to compute the p-values under the null hypothesis of equal performance.

5 Result

Table 2 shows the performance of very negative classification (MN vs. NMN) performed on our data collection. In these experiments, we combine each n-gram model with the rest of features. The n-gram models are unigrams (1g) and unigrams with bigrams (1g 2g), each one weighted with TF-IDF and CountVector. These models were considered as baselines. Then, combined each baseline with one of the rest of features: namely, Doc2vec, SOTF, *VERY-NEG*, (see Table 1). Moreover, we also combined all features with each baseline (All).

In Table 2, we also report the performance of very positive classification (MP vs. NMP) on our dataset. As we did with the most negative classification, n-gram-based classifiers were regarded as baselines, and we examined the association of various combinations of features into the baseline classifiers, including configurations combining all features.

The results depicted by Table 2 show the following trends. Concerning the classification of not very extreme opinions (NMN and NMP), the baseline approaches are already very accurate and, so, the use of the rest of features does not provide any significant improvement. By contrast, the classification of very extreme opinions is a more tough task in which the baselines are outperformed by some of the other features we have tested. The last column in both tables shows the significant differences concerning only MN and MP classifications So, significant tests are shown for classification of extreme opinions. In the case of not extreme opinions, there are no significant improvements when we combine different features.

To detect extreme opinions (both very negative and very positive), the most valuable features are textual features (SOTF) and embeddings (Doc2Vec). However, Doc2Vec is more beneficial to detect the very negative reviews, while SOTF

[6] http://scikit-learn.org/stable/.

Table 2. Polarity classification results, in terms of precision, recall, and F1 scores of (MN Vs. NMN) and (MP Vs. NMP). For each n-gram-based model the best performance for each metric is in bold. The symbol “$\gg$” and “$\ll$” indicates a significant improvement with respect to the n-gram-based baselines, with p-value $\leq$ 0.01. The symbol “$>$” or “$<$” means that the 0.01 $<$ p-value $\leq$ 0.05. “$\sim$” indicate that the difference was not statistically significant (p-value $>$ .05).

Features	MN			NMN			s-test	MP			NMP			s-test
	P	*R*	*F*1	*P*	*R*	*F*1		*P*	*R*	*F*1	*P*	*R*	*F*1	
1g(TF-IDF)	0.75	0.64	0.69	0.89	0.93	0.91		0.87	0.83	0.85	0.94	0.96	0.95	
+ Doc2Vec	0.77	0.70	0.73	0.91	0.93	0.92	$\gg$	0.89	0.85	0.87	0.95	0.96	0.96	$>$
+ SOTF	0.76	0.66	0.71	0.90	0.93	0.91	$>$	0.88	0.87	0.87	0.95	0.96	0.96	$\gg$
+ *VERY-NEG*	0.76	0.65	0.70	0.89	0.93	0.91	$\sim$	0.87	0.83	0.85	0.94	0.96	0.95	$\sim$
+ All	**0.78**	**0.72**	**0.75**	0.91	0.93	0.92	$\gg$	0.89	0.87	0.88	0.96	0.96	0.96	$\gg$
1g(CountVector)	0.67	0.66	0.66	0.89	0.90	0.89		0.81	0.79	0.80	0.93	0.93	0.93	
+ Doc2Vec	0.72	0.70	0.71	0.91	0.91	0.91	$\gg$	0.85	0.84	0.85	0.95	0.95	0.95	$\gg$
+ SOTF	0.68	0.68	0.68	0.90	0.90	0.90	$>$	0.84	0.83	0.83	0.94	0.94	0.94	$\gg$
+ *VERY-NEG*	0.68	0.67	0.67	0.89	0.90	0.90	$\sim$	0.82	0.80	0.81	0.93	0.94	0.94	$>$
+ All	**0.74**	**0.71**	**0.73**	0.91	0.92	0.91	$\gg$	0.86	0.84	0.85	0.94	0.95	0.95	$\gg$
1g 2g(TF-IDF)	0.77	0.62	0.69	0.89	0.94	0.91		0.88	0.84	0.86	0.94	0.96	0.95	
+ Doc2Vec	0.79	0.69	0.74	0.90	0.94	0.92	$\gg$	0.89	0.86	0.87	0.95	0.96	0.96	$\sim$
+ SOTF	0.78	0.64	0.70	0.89	0.94	0.92	$>$	0.88	0.87	0.88	0.96	0.96	0.96	$\gg$
+ *VERY-NEG*	0.68	0.73	0.70	0.89	0.94	0.91	$\sim$	0.88	0.84	0.86	0.95	0.96	0.95	$\sim$
+ All	**0.81**	**0.72**	**0.76**	0.91	0.94	0.93	$\gg$	0.91	0.89	0.90	0.96	0.97	0.97	$\gg$
1g 2g(CountVector)	0.69	0.65	0.67	0.89	0.91	0.90		0.83	0.81	0.82	0.93	0.94	0.94	
+ Doc2Vec	**0.75**	**0.70**	**0.73**	0.91	0.93	0.92	$\gg$	0.86	0.85	0.86	0.95	0.95	0.95	$\gg$
+ SOTF	0.71	0.67	0.69	0.90	0.91	0.90	$\gg$	0.86	0.84	0.85	0.94	0.95	0.95	$\gg$
+ *VERY-NEG*	0.71	0.66	0.68	0.89	0.91	0.90	$\sim$	0.84	0.81	0.82	0.94	0.94	0.94	$\sim$
+ All	0.71	0.68	0.69	0.90	0.91	0.91	$>$	0.89	0.88	0.88	0.96	0.96	0.96	$\gg$

performs better with the very positive ones. Both types of features leads to statistically significant improvements when they are combined with the baselines (n-gram representations). This confirms the valuable information provided by Doc2Vec and SOTF to detect the most extreme reviews. Lexicon-based features slightly improves the baselines but not in a significant way.

Besides, in all cases the combination of all features always yield significant improvements with regard to the baselines. Finally, it is worth noting that none of the features hurts the overall performance.

6 Conclusions

In this article, we have studied different linguistic features for a particular task in Sentiment Analysis. More precisely, we examined the performance of these features within supervised learning methods (using Support Vector Machine (SVM)), to identify extreme opinions on reviews dataset of hotels. The experiments we carried out showed that n-gram models are difficult to outperform, but we found two features that consistently outperforms the baselines: neural-based embeddings and textual features. Polarity lexicons help improve the results, but their influence is moderate. In future work, we will try to compare unsupervised method based to polarity lexicons with the supervised classification described in the current paper.

Acknowledgements. This work has received financial support from TelePares (MINECO, ref:FFI201 4-51978-C2-1-R), and the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

1. Almatarneh, S., Gamallo, P.: Automatic construction of domain-specific sentiment lexicons for polarity classification. In: De la Prieta, F., et al. (eds.) PAAMS 2017. AISC, vol. 619, pp. 175–182. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-61578-3_17
2. Almatarneh, S., Gamallo, P.: Searching for the most negative opinions. In: Różewski, P., Lange, C. (eds.) KESW 2017. CCIS, vol. 786, pp. 14–22. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69548-8_2
3. Almatarneh, S., Gamallo, P.: A lexicon based method to search for extreme opinions. PloS ONE **13**(5), e0197816 (2018)
4. Chenlo, J.M., Losada, D.E.: An empirical study of sentence features for subjectivity and polarity classification. Inf. Sci. **280**, 275–288 (2014)
5. Dai, A.M., Olah, C., Le, Q.V.: Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998 (2015)
6. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Comput. Intell. **22**(2), 110–125 (2006)
7. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014)
8. Lee, J., Park, D.H., Han, I.: The effect of negative online consumer reviews on product attitude: an information processing view. Electron. Commer. Res. Appl. **7**(3), 341–352 (2008)
9. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, Cambridge (2015)
10. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Comput. Intell. **29**(3), 436–465 (2013)
11. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124. Association for Computational Linguistics (2005)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
13. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(Oct), 2825–2830 (2011)
14. Potts, C.: On the negativity of negation. Semant. Linguist. Theory **20**, 636–659 (2010)
15. Potts, C.: Developing adjective scales from user-supplied textual metadata. In: NSF Workshop on Restructuring Adjectives in WordNet, Arlington, VA (2011)

16. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
17. Tripathy, A., Agrawal, A., Rath, S.K.: Classification of sentiment reviews using n-gram machine learning approach. Expert Syst. Appl. **57**, 117–126 (2016)
18. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49. ACM (1999)
19. Zhang, Z., Ye, Q., Zhang, Z., Li, Y.: Sentiment classification of internet restaurant reviews written in Cantonese. Expert Syst. Appl. **38**(6), 7674–7682 (2011)