Searching for the Most Negative Opinions

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 been focused on several aspects of opinions, such as their automatic extraction, identification of their polarity (positive, negative or neutral), the entities or facets involved, and so on. However, to the best of our knowledge, no sentiment analysis approach has considered the automatic identification and extraction of the most negative opinions, in spite of their significant impact in many fields such as industry, trade, political and socials issues.

In this article, we will use diversified linguistic features and supervised machine learning algorithms so as to examine their effectiveness in the process of searching for the most negative opinions.

Keywords: Sentiment analysis · Opinion mining · Linguistic features · Classification · Most negative opinion

1 Introduction

A fundamental task in opinion mining is polarity classification. Polarity classification occurs when a piece of text stating an opinion is classified into a predefined set of polarity categories (e.g., positive, neutral, negative). Categorizing reviews into classes such as "thumbs up" versus "thumbs down," or "like" versus "dislike" are examples of two-class polarity classification [8,9,13,16,17,24–26].

A still not usual way of performing sentiment analysis is to detect and classify the most negative opinions about a topic, object or individual. The most negative opinion is the worst judgment, or appraisal formed in mind about a particular matter. These opinions only constitute a small portion of all opinions found in Social Media. According to [16], only about 5% of all opinions are in the most negative level of the opinion scale, which makes their automatic search a challenge. There is a need for systematic studies that attempt to understand how to mine the vast amounts of unorganized text data and extract the most negative comments.

The objective of the article is to investigate the effectiveness of linguistic features and supervised machine learning classification to search for the most negative opinions. The rest of the paper is organized as follows. In Sect. 2 we discuss the related work. Then, Sect. 3 describes the method. The Experiments are introduced in Sect. 4, where we also describe the evaluation and discuss the results. We draw conclusions in Sect. 5.

P. Różewski and C. Lange (Eds.): KESW 2017, CCIS 786, pp. 14–22, 2017.
https://doi.org/10.1007/978-3-319-69548-8_2

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.

In machine learning techniques there are are two methods, supervised learning, where the most existing techniques for document-level classification use, although there are also unsupervised methods. The success of both mainly depends on the choice and extraction of the proper set of features used to identify sentiments. The current reviews and books in sentiment analysis [1,3,4,14,15,22] included all issues in this field. For instance, the most important linguistic features that used in sentiment classification are listed in Chap. 3 of [15] book. [5] presented a systematic study of different sentence features for two tasks in sentiment classification in (polarity classification and subjectivity classification) our study.

On the other hand, Sentiment words are the core component in opinion mining and have been used in many studies [2,7,11,12,21,23] they relied on lexicons as a source for determining the polarity of documents.

In this study, we focused on searching for the most negative opinions by use linguistic features, because of the vast importance of these views. Previous works analyzed this importance, such as the experiments reported in [6], which found that one-star reviews hurt book sales on Amazon.com. The impact of 1-star reviews that represent the most negative views is higher than the impact of 5-star reviews. [18] also stated that the negative reviews have more impact than positive reviews.

3 The Method

Sentiment analysis typically works at three levels of granularity, namely, document level, sentence level, and aspect level.

Document-level works with whole documents as the basic information unit. Analogously, at the sentence level, sentiment classification is applied to individual sentences in a document. But concerning aspect level, the system performs at a finer-grained level of analysis. Instead of looking at language constructs such as documents, paragraphs, sentences, clauses or phrases, a system working at the aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of, at least, a sentiment (positive, negative or neutral) and a target, namely the aspect of an entity receiving that opinion.

In this paper, however, we are involved with document-level classification issues, more precisely with the identification of most negative opinion *vs.* other opinions at the document level. This binary categorization can be achieved by the use of classifiers built from training data. Converting a portion of text into a feature vector is the essential and basic step in any data-driven approach to Sentiment Analysis. Selection of features is a requirement to make the learning task efficient and accurate. In our experiments, we studied different strategies and examined the following sets of features.

3.1 Unigram Features

First, all stop words are removed from the document collection. Then, the vocabulary is cleaned up by eliminating those terms appearing in less than 12 documents so as to eliminate terms that are too infrequent. Finally, we assign a weight to all terms by using Term Frequency - Inverse Document Frequency (TF-IDF), which is computed in Eq. 1.

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

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.

3.2 Part of Speech Features

A part of speech (PoS) is a category classifying words with similar grammatical properties. PoS tag information is usually used in sentiment analysis and opinion mining. Several researchers [5,8,25] used PoS tags, especially adjectives, as features to classify opinions, such they are a good indicator of sentiment. We processed the document collection using the Natural Language Toolkit (NLTK)[1], which provides words with Penn Treebank PoS tags (see Table 1). Then we counted the occurrences of each tag in the document.

Table 1. Penn Treebank Part-Of-Speech (POS) tags.

CC	conjunction, coordinating	PRP$	pronoun, possessive
CD	cardinal number	RB	adverb
DT	determiner	RBR	adverb, comparative
EX	existential there	RBS	adverb, superlative
FW	foreign word	RP	adverb, particle
IN	conjunction, subordinating or preposition	SYM	symbol
JJ	adjective	TO	infinitival to
JJR	adjective, comparative	UH	interjection
JJS	adjective, superlative	VB	verb, base form
LS	list item marker	VBZ	verb, 3rd person singular present
MD	verb, modal auxillary	VBP	verb, non-3rd person singular present
NN	noun, singular or mass	VBD	verb, past tense
NNS	noun, plural	VBN	verb, past participle
NNP	noun, proper singular	VBG	verb, gerund or present participle
NNPS	noun, proper plural	WDT	wh-determiner
PDT	predeterminer	WP	wh-pronoun, personal
POS	possessive ending	WP$	wh-pronoun, possessive
PRP	pronoun, personal	WRB	wh-adverb

[1] http://www.nltk.org/.

3.3 Syntactic Patterns

We used in this study the patterns defined by Turney [25]. More precisely, he used five patterns of PoS tags to extract opinions from reviews, as the example depicted in Table 2. We define two types of features based on PoS patterns: counting patterns frequency and considering presence or absence of patterns in each document.

Table 2. Pattern of POS by Turney [25]

First word	Second word	Third word
JJ	NN or NNS	Anything
RB, RBR, or RBS	JJ	not NN nor NNS
JJ	JJ	not NN nor NNS
NN or NNS	JJ	not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, or VBG	Anything

3.4 Sentiment Lexicons

In our approach, we have experimented with some lexicons: the Opinion Lexicon or (Sentiment Lexicon), Linguistic Inquiry and Word Count (LIWC) and VADER Lexicon.

- Opinion Lexicon (or Sentiment Lexicon): This is a list of negative and positive sentiment words for English: 5,789 words, 2,006 are positive words and 3,783 are negative. This list has been compiled for many years and its construction was reported in [11]. It includes mis-spellings, morphological variants, slang, and social-media mark-up. The features based on this lexicon are defined by considering the number of negative and positive terms in the document, as well as the proportion of negative and positive terms.
- Linguistic Inquiry and Word Count (LIWC): [21] LIWC dictionary consists of 290 words and word-stems. Each word or word-stem defines one or more word categories or sub-dictionaries. We believe that the use of features derived from the LIWC dictionary (Linguistic Inquiry and Word Count) would be helpful in the search for the most negative opinions since negative opinions can also be associated with psychological factors. We obtained 65 features based on the lexical categories defined in LIWC.
- Valence Aware Dictionary and Sentiment Reasoner (VADER): This is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains [12]. We obtained over 7,500 lexical features with validated valence scores indicating both sentiment polarity (negative/positive) and sentiment intensity on a scale from -4 to $+4$. Intensity was classified as follows.

Words were split into four groups according to valence scores: −4 to −2 most negative, −1.9 to −0.1 negative, +0.1 to +1.9 positive and +2 to +4 most positive. Then the number and proportion of each group of words were considered to define the intensity-based features. Also, we included additional features: namely, the total scores for all the words that appear in the documents and the total scores of words that are only provided with negative scores in the documents.

4 Experiments

4.1 Data Collection

In order to extract the most negative opinions, we require to analyze document collections with scaled opinion levels (e.g. rating) and extract those documents associated with the lowest scale. So, we have adopted (Pang & Lee Sentiment scale dataset)[2], which was described in [16]. This dataset contains four corpora of movie reviews, where each corpus includes documents written by the same author. The total number of documents in all corpus are 5,006.

4.2 Training Set

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 [17]. We decided to utilize *scikit*[3] which is an open source machine learning library for the Python programming language [20]. 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. More specifically, we utilized the sklearn.svm.LinearSVC module[4]. Our collection has 5,006 reviews and our method handles a large number of features for each example. To do classification, we need two samples of documents: training and testing. The training sample will be used to learn various characteristics of the documents and the testing sample was used to predict and next verify the efficiency of our classifier in the prediction. So we divided the dataset into two stratified samples: we have allocated 25% of the collection for the testing sample and 75% of the collection for the training sample.

There are only 615 most negative reviews out of 5,006 in our dataset and 4,394 labeled as a negative class (not most negative), which results in an unbalanced two-class classification problem. To deal with this problem there are many frameworks and approaches such as undersampling and oversampling, even if undersampling gives rise to loss of information. As recommended in [10,19], we examined the performance by giving more importance to the positive class.

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/.
[3] http://scikit-learn.org/stable/.
[4] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html.

We found that performance was insensitive to the SVM cost parameter (C) but very sensitive to the weights that modify the relative cost of misclassifying positive and negative samples.

In our analysis, we employed 5_fold cross_validation and the effort was put on optimizing F1 which is computed with respect to the most negative opinions (which is the target class):

$$F1 = 2 * \frac{P * R}{P + R} \tag{2}$$

where P and R are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

where TP stands for true positive, FP is false positive, and FN is false negative.

To optimize F1, we tried out a grid search approach with exponentially growing sequences of the value of the parameter *class_weight*. More precisely, we tested class_weight with different values: $2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, ..., 2^{10}$. After finding the best value of class_weight within that sequence, we conducted a finer grid search on that better district (e.g. if the optimal value of class_weight is 8, then we test all the neighbors in this region: e.g. 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15 and 16).

The class_weight was finally set to the value returning the highest F1 across all these experiments (see Table 3).

Table 3. The best (F1) performance with varying class weights

Features	Class-weight
Unigram (TF-IDF)	893
Unigram (Pres)	2
POS	4
Pattern (Freq)	8
Pattern (Presence)	8
Opinion Lexicon	6
LIWC	6
VADER	6
ALL	4

Figure 1 shows the average of F1 performance across the variation of class_weight for each set of features.

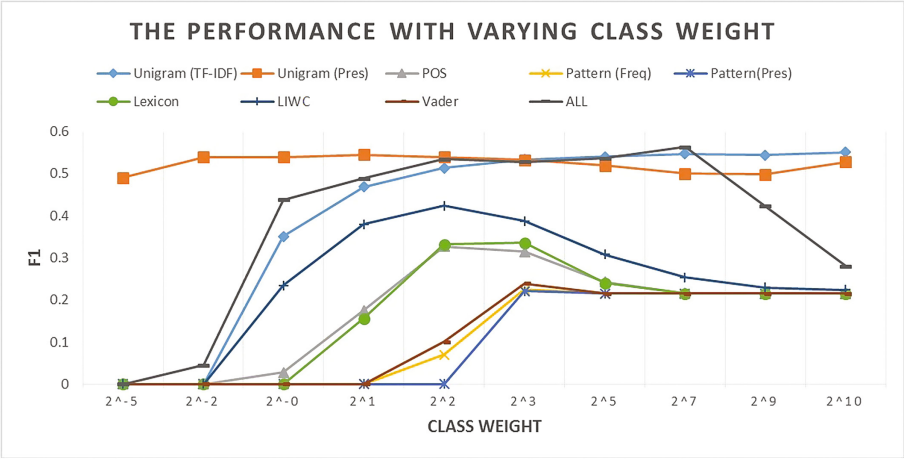

Fig. 1. The average performance of F1 with across different values of class_weight

4.3 The Results

In the test collection, there are 1,252 reviews and 157 of them belong to the target class (the most negative opinions). The proportion of positive examples in the training and test collections are similar (around 12%); consequently, both datasets are similarly unbalanced. The results depicted in Table 4 reveal that all combined features give the best performance in terms of precision and F1, even though just unigrams work reasonably well.

In order to select the best and most influential singular features for finding most negative opinions, we need to perform further fine-grained experiments with different groups of feature combinations.

Table 4. The best results for the collection, in terms of precision, recall, and F1 scores

Features	Precision	Recall	F1
Unigram (TF-IDF)	0.60	0.54	0.57
Unigram (Pres)	0.63	0.47	0.54
POS	0.25	0.33	0.29
Pattern (Freq)	0.14	0.73	0.24
Pattern (Presence)	0.13	0.71	0.22
Opinion Lexicon	0.25	0.61	0.36
LIWC	0.30	0.62	0.40
VADER	0.18	0.29	0.22
ALL	0.69	0.54	0.61

5 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 the most negative documents on movie review datasets.

The experiments reported in our work shows that the evaluation values for identifying the most negative class are low. This can be partially explained by the difficulty of the task, since the difference between very negative and not very negative is a subjective continuum without clearly defined edges. The borderline between very negative and not very negative is still more difficult to find than that discriminating between positive and negative opinions, since there are a quite clear space of neutral/objective sentiments between the two opinions. However, there is not such an intermediate space between *very* and *not very*.

In future work, there is much room for improvement. First, use more data sets such as products reviews as well as movie reviews. Second, we will provide the classifiers with a set of features which would be sensitive to the concept of *most negative*. Third, it would be useful to make experiments with unsupervised learning approaches and lexicon-based methods to improve the performance for this difficult task.

References

1. Agarwal, B., Mittal, N.: Prominent Feature Extraction for Sentiment Analysis. Socio-Affective Computing. Springer, Cham (2016). doi:10.1007/978-3-319-25343-5
2. Almatarneh, S., Gamallo, P.: Automatic construction of domain-specific sentiment lexicons for polarity classification. In: De la Prieta, F., Vale, Z., Antunes, L., Pinto, T., Campbell, A.T., Julián, V., Neves, A.J.R., Moreno, M.N. (eds.) PAAMS 2017. AISC, vol. 619, pp. 175–182. Springer, Cham (2018). doi:10.1007/978-3-319-61578-3_17
3. Benamara, F., Taboada, M., Mathieu, Y.: Evaluative language beyond bags of words: linguistic insights and computational applications. Comput. Linguist. **43**, 201–264 (2017)
4. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**(2), 15–21 (2013)
5. Chenlo, J.M., Losada, D.E.: An empirical study of sentence features for subjectivity and polarity classification. Inf. Sci. **280**, 275–288 (2014)
6. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: online book reviews. J. Mark. Res. **43**(3), 345–354 (2006)
7. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 231–240. ACM (2008)
8. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, pp. 174–181. Association for Computational Linguistics (1997)

9. Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., de Jong, F.: Polarity analysis of texts using discourse structure. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1061–1070. ACM (2011)
10. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
11. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
12. Hutto, C.J., Gilbert, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
13. Kamps, J., Marx, M., Mokken, R.J., De Rijke, M., et al.: Using wordnet to measure semantic orientations of adjectives. In: LREC, vol. 4, pp. 1115–1118 (2004)
14. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
15. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, Cambridge (2015)
16. 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)
17. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
18. Papathanassis, A., Knolle, F.: Exploring the adoption and processing of online holiday reviews: a grounded theory approach. Tourism Manag. **32**(2), 215–224 (2011)
19. Parapar, J., Losada, D.E., Barreiro, A.: A learning-based approach for the identification of sexual predators in chat logs. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
21. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. Technical report (2015)
22. Serrano-Guerrero, J., Olivas, J.A., Romero, F.P., Herrera-Viedma, E.: Sentiment analysis: a review and comparative analysis of web services. Inf. Sci. **311**, 18–38 (2015)
23. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
24. Takamura, H., Inui, T., Okumura, M.: Extracting semantic orientations of phrases from dictionary. In: HLT-NAACL, vol. 2007, pp. 292–299 (2007)
25. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)
26. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 129–136. Association for Computational Linguistics (2003)

http://www.springer.com/978-3-319-69547-1

Knowledge Engineering and Semantic Web
8th International Conference, KESW 2017, Szczecin, Poland, November 8-10, 2017, Proceedings
Różewski, P.; Lange, C. (Eds.)
2017, XIII, 364 p. 92 illus., Softcover
ISBN: 978-3-319-69547-1