Language-Independent Sentiment Polarity Detection in Movie Reviews: A Case Study of English and Spanish

Jelena Graovac and Gordana Pavlović-Lažetić

University of Belgrade, Faculty of Mathematics, Department of Computer Science,
Studentski trg 16, 11000 Belgrade, Serbia
{jgraovac,gordana}@matf.bg.ac.rs

Abstract. We present a novel language-independent technique for determining polarity, positive or negative, of opinions expressed by different individuals. The technique is based on byte-level n-gram frequency statistics method for document representation, and a variant of k nearest neighbors (kNN) (for $k = 1$) machine learning algorithm for categorization process. The main advantages of the technique are its simplicity and full language and topic independence. For driving experiments we used corpora of movie reviews: Cornell polarity dataset in English and MuchoCine in Spanish. Experimental results (85.6% accuracy for English and 82.49% for Spanish corpora) confirm that the presented technique is comparable with the best ranked previously published techniques, when applied to movie reviews datasets. Still, it use no additional linguistic information nor external resources.

Keywords: Sentiment Analysis, Byte n-Grams, kNN, Movie Review.

1 Introduction

Sentiment Analysis (SA) is a challenging task that combines natural language processing and text mining techniques in order to automatically identify and analyze opinions and emotions in documents. This relatively new area of research is becoming more and more important mainly due to the explosion of the Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media. Internet has turned into a collaborative framework where social and technological trends come together [11]. One of the most popular tasks related to SA is Sentiment Polarity Detection (SPD), which focuses on determining the overall sentiment-orientation (positive or negative) of the opinions expressed by individuals. SPD has attracted a great deal of attention, in part because of its potential applications. It has proven useful for companies, recommender systems, and editorial sites to create summaries of people's opinions and experiences that consist of subjective expressions extracted from review's polarity – positive or negative [17]. It could also be useful in text summarization, message filtering and many other business intelligence applications.

Different approaches have been used for SPD, but the mainstream basically consists of two major methodologies: A supervised Machine Learning (ML) methodology based on using a collection of data to train the classifiers [17] and unsupervised methodology based on a semantic orientation applied when linguistic resources are available [23]. In order to take advantage of both methodologies, some studies apply a hybrid approach. Regardless of which approach we choose, we are faced with a number of challenges to deal with. Pang and Lee [16] concluded that the sentiment classification problem is more challenging than traditional topic-based categorization problem. While topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner. They give an example of a sentence in a movie review: "How could anyone sit through this movie", that contains no single word that is obviously negative. Thus, they conclude that sentiment seems to require more understanding than ordinary topic-based classification. A common phenomenon in movie reviews is a kind of "thwarted expectations" narration, where the author deliberately gives the opposite conclusion compared to the previous discussion, which further complicates the SA. Other difficulty in handling texts written by web users is the presence of different kinds of textual errors, such as typing, spelling and grammatical errors. Also, although most of the research activity on sentiment analysis has concentrated on English text, people increasingly comment on their points of views, experiences and opinions in many other languages. Using different languages produces additional difficulty in SA regarding specific features of the languages. The management and study of SA in languages other than English is a growing need.

The aim of this paper is to present a byte-n-gram-based language independent technique that has been successfully used in solving topic-based text categorization task [6] and to apply it to the task of SPD in movie reviews, avoiding many difficulties listed above. Note that the Turney [23] found movie reviews to be the most difficult of several domains for sentiment classification. Since English and Spanish are among top three languages most used in the Internet according to the Internet World State rank[1], we focus on English and Spanish SPD. For driving experiments, we use the following corpora of movie reviews: Cornell Polarity Dataset [17] in English and MuchoCine [3] in Spanish.

The rest of the paper is organized as follows: the next section presents previous related work on SPD, regarding English and non-English texts. Section 3 presents the technique proposed in our work and Section 4 describes the experimental framework. The results obtained by experiments are expounded in Section 5, as well as comparisons with previously published results obtained over the same datasets that we use in this work. Section 6 concludes the paper.

[1] http://www.internetworldstats.com/stats7.htm

2 Related Work

In this section, a particular attention is given to references that discuss the SPD problem using the same corpora that we use in our paper. The corresponding results will be used in our comparative study.

Sentiment Analysis in English. Supervised ML algorithms such as Support Vector Machines (SVM), Maximum Entropy (ME) and Naive Bayes (NB) were used by Pang and Lee ([16], [17]) to classify movie reviews in English. Authors used bag-of-words representation of documents ignoring word order and syntactic relations between words. This information was incorporated into document sentiment classification by Matsumoto and his colleges in [14], improving accuracy of ML algorithms. In [12] Martineau and Finin introduced new "delta tf-idf" weight (that places a much greater weight on sentimental words) that improved SVM algorithm for classification. Matsumoto and others in [14] achieve very good results using language-independent feature weights, but results were better when they used additional English specific linguistic information. In [19] Raychev and Nakov used language-independent weights assigned to words and word bigrams for document representation and Naive Bayes classifier. They achieved very good results by using subjectivity dataset.

Much research in English has been at least partially knowledge-based ([23]). In [7] Kennedy and Inkpen presented two methods for determining the sentiment expressed by a movie review: unsupervised (they examined the effect of valence shifters on classifying the reviews) and supervised SVM method. Hybrid approach was applied by Konig and Brill in [10]. Authors constructed a hybrid classifier that utilizes human reasoning over automatically discovered text patterns to complement machine learning. In [18] Prabowo and Thelwall combined Rule-Based Classification (RBC), supervised learning and machine learning into a new combined method. In [25] Whitelaw and his colleges presented a new hybrid method showing that useing features based on appraisal group analysis can significantly improve sentiment classification. Wang and Domeniconi [24] embedded background knowledge derived from Wikipedia into a semantic kernel, used to enrich the BOW representation of documents and to improve SVM classification.

Sentiment Analysis in non-English Languages. There are some interesting papers that have studied the problem using non-English collections including German ([9]), Chinese ([22]), French ([1]), or Arabic ([20]). Regarding SA focused on Spanish, the MuchoCine corpus used in this work has been widely used. Martinez and colleagues [13] applied the supervised approach to this corpus using different machine learning algorithms (SVM, NB, BBR, KNN, C4.5). Del-Hoyo [4] and others defined a hybrid statistical-semantic system for opinion detection in Spanish language texts. Martin-Valdivia with others [11] also presented hybrid approach of supervised (using SVM, NB, C4.5, BBR) and unsupervised (using SentiWordNet) methods. They obtained much better results than those obtained with the unsupervised approach proposed by Cruz [3] and Molina-González [15].

3 Byte-n-Gram-Based Classification Technique

The technique for SPD that we propose in this paper is based on a byte-level n-gram frequency statistics method for document representation, derived from Kešelj's n-gram based method for authorship attribution [8], and a variant of kNN (for $k = 1$) machine learning algorithm for categorization process. The term n-gram could be defined on a word, character or byte level. Extracting byte n-grams from a document is like moving an n-byte wide "window" across the document, byte by byte. Each window position covers n bytes, defining a single n-gram. In the case of Latin-alphabet languages, character-level and byte-level n-gram models are quite similar according to the fact that one character is usually represented by one byte. The only difference is that character-level n-grams use letters only and typically ignore digits, punctuation, and whitespace while byte-level n-grams use all printing and non-printing characters. Since our technique is based on byte level n-grams, it has a lot of advantages: language and topic independence, relative insensitivity to spelling variations/errors, word stemming is got essentially for free, no linguistic knowledge is required, independence of encoding and alphabet, only one pass processing is required (for more details see [6]). The main disadvantage of using n-grams is that it yield a large number of n-grams.

3.1 Categorization Procedure

The categorization procedure is divided into two stages: *Training stage* (construct sentiment classifier) and *Testing stage* (classify movie reviews).

Training Stage. For a given training data of movie reviews divided into two categories – positives and negatives, build a sentiment classifier:

- Concatenate all the training documents that belong to the same category into a single document. Each category will be thereby presented by one document only.
- For each category document and test document, construct its profile:
    - Select a specific n-gram size n (e.g. 9-gram, 10-gram etc.).
    - Extract the byte-level n-grams for that particular value of n and calculate the normalized (relative) frequencies, for each n-gram.
    - List the n-grams by descending frequency, so that the most frequent are listed first.
    - Select a specific profile length L at which to cut off all test document and category profiles.

Testing Stage. For given test data, assign each test document (movie review) one or more categories (positive or negative):

- Compute a dissimilarity measure between the test document's profile and each of the category's profiles.

– Select the category (or categories) whose profile has the smallest value of dissimilarity measure with the document's profile.

Following this procedure, a category profile and a test document profile will be simply a set of L pairs $\{(x_1, f_1), (x_2, f_2)...(x_L, f_L)\}$ of the most frequent n-grams and their normalized frequencies. In order to decide whether a certain test document belongs (or not) to a certain category, this text categorization procedure requires a dissimilarity measure.

Dissimilarity Measures. In this paper we use three dissimilarity measures. The first one is the original dissimilarity measure used by Kešelj [8] and it has a form of relative distance:

$$dK(\mathcal{P}_1, \mathcal{P}_2) = \sum_{n \in profile} \left(\frac{2 \cdot (f_1(n) - f_2(n))}{f_1(n) + f_2(n)} \right)^2 \quad (1)$$

where $f_1(n)$ and $f_2(n)$ are frequencies of an n-gram n in the category profile P_1 and the test document profile P_2, respectively.

The second measure is introduced by Cavnar and Trenkle [2] and it is a simple rank-order statistic. For each n-gram in a test document's profile, its counterpart in a category's profile is located, and then calculated how far out of place it is. If an n-gram is not in the category's profile, it takes a maximum out-of-place value, which is equal to the number of n-grams in the profile. The sum of all of the out-of-place values for all n-grams is the dissimilarity measure between the document and the category profiles. We will refer to this measure as dOP (Out-of-Place).

The last dissimilarity measure used in this paper is introduced by the first author of this paper in [6]. It represents the number of n-grams that appear in the union of the profiles and not in their intersection. In mathematics, this is known as symmetric difference, so we will refer to this measure as dSD:

$$dSD(P_1, P_2) = |P_1 \triangle P_2| \quad (2)$$

where P_1 is a category profile and P_2 is a test document profile.

Implementation Details. For producing n-grams and their normalized frequencies, the software package *Ngrams* written by Kešelj [8] is used. For the process of categorization, the software package *NgramsCategorization* developed by the first author of this paper is used. Source code can be obtained on request.

4 Experimental Framework

4.1 Evaluation Measures

We have used the typical evaluation measures used in text classification: Precision (P), Recall (R), Accuracy (Acc) and F1:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, Acc = \frac{TP + TN}{TP + TN + FP + FN}, F1 = \frac{2PR}{P + R} \quad (3)$$

where TP (True Positives) is defined as the number of documents that were correctly assigned to the considered category, TN (True Negatives) is the number of the assessments where the system and a human expert agree on a negative label, the FP (False Positives) is the number of positive labels assigned by the system that do not agree with the expert assignment, and FN (False Negatives) is the number of negative labels that the system assigned to documents otherwise assessed as positive by the human expert [21].

All presented measures can be aggregated over all categories (positive and negative) in two ways: micro-averaging – the global calculation of measure considering all the documents as a single dataset regardless of categories, and macro-averaging – the average on measure scores of all the categories.

4.2 Data Collections

For our experiments, we chose to work with movie reviews in English and Spanish.

Cornell Polarity Dataset (CPD) - in English. This corpus[2] is firstly introduced by Pang and Lee [17]. It contains 1000 positive and 1000 negative reviews (we used here version 2.0) and it is compiled before 2002, with 20 reviews per author (312 authors total) per category. Testing and training data are randomly distributed in the ratio 2 : 1.

MuchoCine (MC) - in Spanish. The corpus [3] consists of 3878 movie reviews collected from the MC website. For this study, ”neutral” opinions (movies with a score of 3 out of 5) have not been used, so the total number of documents on which the experiments have been performed is 2625, with 1274 negative reviews, and 1351 positive reviews. Testing and training data are randomly distributed in the ratio 2 : 1.

5 Experiments and Results

The effectiveness of the technique presented in this paper can be controlled by the two parameters: n-gram size n, and profile length L. The most important question is: What are the values of n-gram size n and profile length L that produce the best accuracy? To give an answer to this question, an extensive set of experiments were conducted over the CPD and MC movie review's corpora. The results for n-gram size n between 6 and 13 and profile length L from 5000 to 100000 with step 5000, in term of macro-average Accuracy, are presented in the upper part of the Fig. 1. For n and L out of these scopes, weaker results are obtained. We conclude that the accuracy peaks at the n-gram size $n = 8$ and $n = 10$ in the case of English CPD, and $n = 11$ in the case of Spanish MC corpus. For these values of n, comparisons between different dissimilarity measures are performed. Results are presented in the bottom part of the Fig. 1. We conclude that dK slightly outperforms other measures. The best accuracy

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/

that we obtained is 82.49% for Spanish MC corpus ($n = 11$, $L = 40000$) and 85.6% for English CPD ($n = 8$, $L = 45000$; $n = 10$, $L = 90000$).

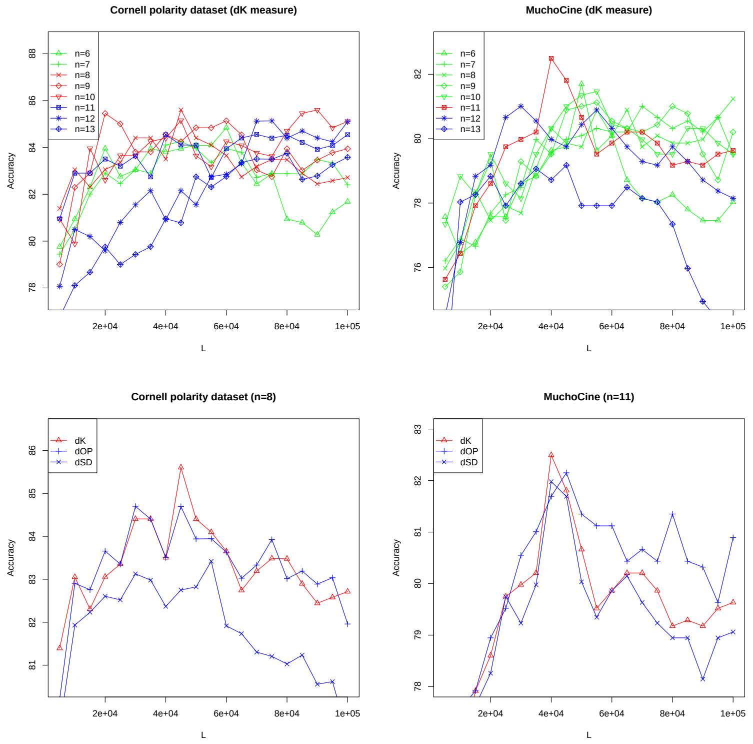

Fig. 1. Accuracy of byte-n-gram-based technique for the English CPD and the Spanish MC corpora of movie reviews. Results are presented for different n-gram size n and dissimilarity measure dK (the upper part of the picture), and different dissimilarity measures with the chosen value of n-gram size n (the bottom part of the picture).

5.1 Comparison With Other Related Work

In order to evaluate performance of the technique presented in this paper, we compare the results with the published results obtained by other methods (briefly presented in Section 2) over the CPD and MC corpora. Comparison results are presented in Tables 1 and 2. Only the best reported results are presented for each technique on each corpus. Because of the diversity of the evaluation methods and different methodologies on which the techniques are based (supervised, unsupervised, hybrid), we need to be cautious in interpreting the results listed in Tables 1 and 2.

Table 1. Polarity classification results in percentages over Cornell polarity dataset in English.

	Technique	Version	Precision	Recall	F1	Accuracy
Pang et al. (2002)[16]	NB, ME, SVM – Supervised	v1.0	N/A	N/A	N/A	77-82.9
Pang et al. (2004)[17]	NB, SVM – Supervised	v2.0	N/A	N/A	N/A	86.4-87.2
Matsumoto et al. (2005)[14]	SVM – Supervised	v2.0	N/A	N/A	N/A	88.1
	SVM – Hybrid	v2.0	N/A	N/A	N/A	92.9
Whitelaw et al (2005)[25]	SVM – Hybrid	v2.0	N/A	N/A	N/A	90.2
Konig and Brill (2006)[10]	SVM – Hybrid	v2.0	N/A	N/A	N/A	91
Kennedy and Inkpen (2006)[7]	Unsupervised	v2.0	68.2	67.8	68	67.8
	SVM – Supervised	v2.0	86.1	86.15	86.15	86.2
Wang and Domeniconi (2008)[24]	SVM - Supervised	v2.0	81.24	N/A	N/A	N/A
	SVM – Hybrid	v2.0	86.37	N/A	N/A	N/A
Rudy and Thelwall (2009)[18]	RBC, SVM – Hybrid	v2.0	N/A	N/A	N/A	83.33-87.29
Martineau and Finin (2009)[12]	SVM – Supervised	v2.0	N/A	N/A	N/A	88.1
Raycev and Nakov (2009)[19]	NB – Supervised	v2.0	N/A	N/A	N/A	89.85
Our proposal	kNN-Supervised	v2.0	**85.64**	**85.61**	**85.6**	**85.6**

Table 2. Polarity classification results over MuchoCine corpus in Spanish.

	Technique	Precision	Recall	F1	Accuracy
Cruz et al. (2008) [3]	Unsupervised	N/A	N/A	N/A	69.5
del-Hoyo et al. (2009) [4]	NN, SVM – Hybrid	N/A	N/A	N/A	79.31-80.86
	NN, SVM – Supervised	N/A	N/A	N/A	77.05-77.13
	NN, SVM – Unsupervised	N/A	N/A	N/A	67.31-67.64
Malvar-Fernández et al. (2011) [5]	SVM – Supervised	77	77	N/A	N/A
Martnez-Cámara et al. (2011) [13]	SVM, NB, BBR, kNN – Supervised	68.15-87.21	68.20-87.01	68.17-87.10	68.13-87.08
Martn-Valdivia et al. (2013) [11]	SVM, NB, C4.5, BBR – Hybrid	87.71-88.58	87.64-88.57	87.66-88.56	87.66-88.57
Molina-González et al. (2013) [15]	Unsupervised	63.93	62.74	63.33	63.16
Our proposal	kNN-Supervised	**83.06**	**82.29**	**82.34**	**82.49**

6 Conclusion and Future Work

In this paper we presented a language-independent byte-n-gram-based technique for polarity classification over a corpus of movie reviews written in English, the Cornell polarity dataset (CPD) and Spanish, the MuchoCine (MC) corpus. The technique is based on byte-level n-gram frequency statistics method for document representation and a variant of k nearest neighbors (for $k = 1$) machine learning algorithm for categorization process. It is simple to use and it is fully language and topic independent, so it can be applied to corpora in other languages and domains, without any changes. Experimental results are promising (82.5% for MC and 85.6% for CPD), comparable with the best ranked published results.

There are many ways in which the presented technique could be extended. First, we shall try to improve it by adding n-gram weighting factors schema, that comes from inter-document source. Another possible research direction would be combining our approach with some language-specific approaches in order to improve accuracy. Also, we plan to apply our technique to other domains and languages.

References

1. Balahur, Alexandra, and Marco Turchi: Multilingual sentiment analysis using machine translation? Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics (2012)
2. Cavnar, William B., and John M. Trenkle: N-gram-based text categorization. Ann Arbor MI 48113.2: 161–175 (1994)
3. Cruz, Fermin L., et al.: Clasificación de documentos basada en la opinión: experimentos con un corpus de criticas de cine en espanol. Procesamiento de Lenguaje Natural 41 (2008)
4. del-Hoyo, Rafael, et al.: Hybrid text affect sensing system for emotional language analysis. Proceedings of the international workshop on affective-aware virtual agents and social robots. ACM (2009)
5. Fernández, Paulo Malvar, and José Ramom Pichel Campos: Generación semiautomática de recursos de Opinion Mining para el gallego a partir del portugués y el español (2011)
6. Graovac, Jelena: A variant of n-gram based language-independent text categorization, Intelligent Data Analysis, Vol. 18, No. 4 (to appear in 2014)
7. Kennedy, Alistair, and Diana Inkpen: Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence 22.2: 110–125 (2006)
8. Kešelj, Vlado and Peng, Fuchun and Cercone, Nick and Thomas, Calvin: N-gram-based author profiles for authorship attribution. Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING, Vol. 3, pp.255–264, (2003)
9. Kim, Soo-Min, and Eduard Hovy: Identifying and analyzing judgment opinions. Proceedings of the main conference on Human Language Technology Conference of the North American. Association for Computational Linguistics (2006)

10. Konig, Arnd Christian, and Eric Brill: Reducing the human overhead in text categorization. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2006)
11. Martín-Valdivia, María-Teresa, et al.: Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications 40.10: 3934–3942 (2013)
12. Martineau, Justin, and Tim Finin: Delta TFIDF: An Improved Feature Space for Sentiment Analysis. ICWSM (2009)
13. Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña-López, L. A.: Opinion classification techniques applied to a Spanish corpus. In Proceedings of the 16th international conference on Natural language processing and information systems, NLDB'11. 169-176. Springer-Verlag (2011)
14. Matsumoto, Shotaro, Hiroya Takamura, and Manabu Okumura: Sentiment classification using word sub-sequences and dependency sub-trees. Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 301–311 (2005)
15. Molina-González, M. Dolores, et al.: Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications 40.18, 7250–7257 (2013)
16. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan: Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics (2002)
17. Pang, Bo, and Lillian Lee: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd annual meeting on Association for Computational Linguistics (2004)
18. Prabowo, Rudy, and Mike Thelwall: Sentiment analysis: A combined approach. Journal of Informetrics 3.2: 143–157 (2009)
19. Raychev, Veselin, and Preslav Nakov: Language-independent sentiment analysis using subjectivity and positional information. Proceedings of the international conference RANLP (2009)
20. Rushdi-Saleh, Mohammed, et al.: OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology 62.10: 2045–2054 (2011)
21. Sebastiani, Fabrizio: Machine learning in automated text categorization, ACM computing surveys (CSUR), Vol. 34, No. 1: 1–47 (2002)
22. Tan, Songbo, and Jin Zhang: An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications 34.4: 2622–2629 (2008)
23. Turney, Peter D: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics (2002)
24. Wang, Pu, and Carlotta Domeniconi: Building semantic kernels for text classification using wikipedia. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM (2008).
25. Whitelaw, Casey, Navendu Garg, and Shlomo Argamon: Using appraisal groups for sentiment analysis. Proceedings of the 14th ACM International Conference on Information and Knowledge Management. ACM (2005)