ISSN 1870-4069

A Lexical Supervised Approach for Opinion Mining in the Domain of Laptops and Restaurants

Karen Vazquez[1], Mireya Tovar[1], David Pinto[1], José A. Reyes-Ortiz[2]

[1] Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla,
Mexico

[2] Autonomous Metropolitan University, Systems Department, Azcapotzalco,
Mexico City, Mexico

krnlet@gmail.com, {mtovar,dpinto}@cs.buap.mx, jaro@correo.azc.uam.mx

Abstract. This paper presents a study of opinion mining or sentiment analysis for detection of polarity in a set of users opinion about restaurants written in Spanish and English. The research work is performed with the aim of solving a task proposed in SemEval 2016, thus we employed the same dataset proposed in that evaluation conference. The proposed approach uses a vector model for representing the information, including lexical features such as the following ones: word unigrams, bigrams and trigrams. The obtained results show a performance up to 71% when using word unigrams for representing the opinions written in English in the domain of restaurants.

Keywords: Opinion mining, vector space model, natural language processing.

1 Introduction

Nowadays, the major of people which is connected to Internet do it through social networks. Social communication media are used as a space for consuming and producing information. Thus, there is a great opportunity for studying, among other things, public opinions of consumers with the aim of providing information to final users and business owners about the quality of the service of, for example, restaurants and computer selling shops. In this manner, it would be possible to know the quality of a given restaurant according to the consumer rankings, or the best place to buy computers (for example, laptops) according to the consumer opinions.

This task has been proposed by SemEval 2016[1], a semantic evaluation forum, among other 13 tasks associated with semantic issues of natural language under-

[1] http://alt.qcri.org/semeval2016/

standing. The task number 5 (aspect-based sentiment analysis), in particular, its subtask 2 (text-level aspect-based sentiment analysis) is the one that has been considered for the experiments carried out in this paper [7]. The final aim is to automatically obtain the "category" (determined by the tuple: aspect-polarity) for a number of opinions about a entity or domain (in this case, restaurants or laptops) given by consumers (users or clients). Thus, the idea is to automatically detect the polarity of those opinions as positive, negative, neutral or conflict.

As contribution to solving this particular task, we propose to employ a vector space model with a number of lexical features based on word unigrams, bigrams and trigrams. The text representation schema considers the use of term frequency (TF) and inverse document frequency (IDF) for obtaining the representative vectors (TF-IDF) taking into account a training dataset provided by SemEval. The obtained results show a good performance when this model is employed.

The remaining of this paper is structured as follows. In Section 2 we present the related work. Section 3 describes the algorithm or model proposed. In Section 4.1 we describe the dataset employed in the experiments. The obtained results are given in Section 4.2. Finally, in Section 5 the conclusions are given.

2 Related Work

In recent years, various approaches have been proposed tackling the task of sentiment analysis through Natural Language Processing. Even if most of the works reported in literature deal with documents written in English, other languages such as Spanish have also been reported. In this section we present some research works related with the topic of this paper.

In [8] it is described the opinion mining system named "sentiue", which claims to determine the polarity of the sentiment expressed about a certain aspect of a target entity. This system participated in the Task 12 of SemEval-2015 obtaining a 79% of accuracy when determining the sentiment polarity of a given text.

In [2] it is presented a contribution to the Task 5 of SemEval 2016, working with documents written in English and French for user opinions for the domain of Restaurants. This system is based on composite models, combining linguistic features with machine learning algorithms. According to the reported results, they obtained 88% of accuracy for determining polarity in the restaurant domain (English language), and 78% of accuracy for determining polarity in the restaurant domain (French language).

In [4], authors describe the system they used in the task 5 of SemEval 2016. Their system is based on supervised machine learning, using a Maximum Entropy classifier, conditional random fields, and a large number of features such as global vectors, Latent Dirichlet Allocation, bag of words, emoticons, and others. They obtained very competitive results in the SemEval competition by using this system.

In [3], it is proposed a supervised term weighting scheme based on two factors: importance of a term in a document (ITD) and importance of a term for expressing sentiment (ITS). For ITD, they explore three definitions based

on term frequency, and seven statistical functions are employed to learn the ITS of each term from training documents with manually annotated categories. The experimental results show that their method produces the best accuracy on two of three data sets.

The main objectives of the approach proposed by [6] are two-fold, first to improve feature-based opinion mining by employing ontologies in the selection of features, and second, to provide a method for sentiment analysis based on vector analysis. Their approach achieved an accuracy of 89.6% for the sentiment classification of the opinions in one of the following classes: positive, negative and neutral.

In the research work conducted by [1], they present an approach based on ontologies matching for opinion analysis. The aim of their work is to allow two enterprises to share and merge the results of opinion analyses on their own products and services.

Martínez Camara et al. [5] tested two classification algorithms (SVM, Naïve Bayes) and several weighting schemes and linguistic preprocessing (stopwords removing and stemmer) to determinate the opinion polarity in the domain of movies using Spanish language. The authors conclude that SVM works better than Naïve Bayes.

As we mentioned before, there are many works reported in literature associated with opinion mining, so we will avoid to be exhaustive on mentioning all of these works and we will proceed to describe the approach employed in our experiments.

3 Description of the approach employed

In this paper we propose an approach for determining the polarity of user opinions provided by organizers of Task 5 of SemEval-2016 [7]. First of all, we apply a preprocessing step to the training data in order to obtain the representative vectors for each tuple {aspect, polarity}. We do exactly the same process for the test dataset so that we can be able to apply the cosine similarity between these two datasets in order to determine the polarity of each element of the test dataset. The performance of the approach is obtained by comparing the results with those reported in the gold standard. Fig. 1 shows, graphically, the algorithm proposed.

PHASE 1

- Preprocessing:
  - Extraction of opinions from the XML document: To filter in order to obtain only the opinions from the XML document.
  - Cleaning of opinions: To remove stop words, punctuation symbols, isolated character and sorting of terms.
  - Tokenization: Tokenize opinions by words in order to obtain the vocabulary of the dataset.
  - Stemming: The aim is to reduce the vocabulary by stemming each word in order to reduce them to a common base form.

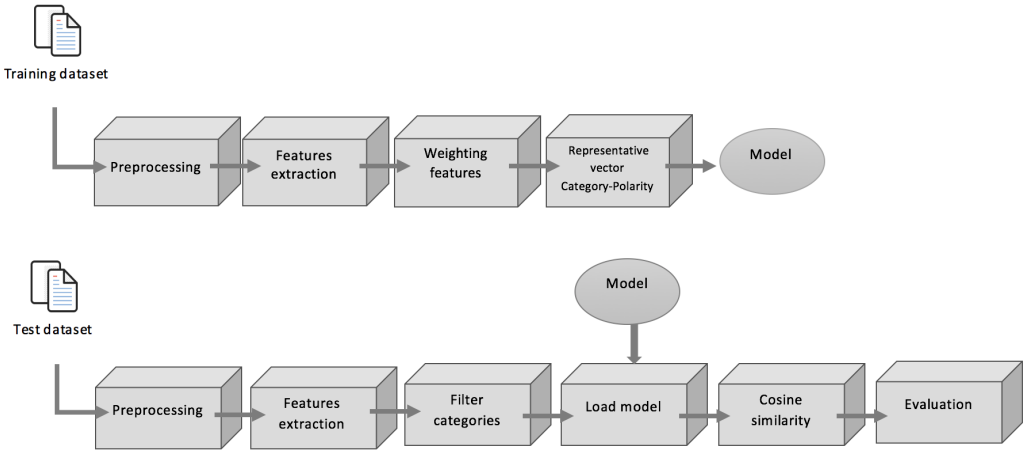

Fig. 1. Proposed algorithm.

  - Filtering categories: To filter opinions of the training dataset by entity and attribute, for example, RESTAURANT#GENERAL, FOOD#PRICES, etc.
- Weighting features extraction:
  - *Term Frequency* (TF): The number of times that a given term appears in a document or dataset, which allows to represent it.
  - *Inverse Document Frequency* (IDF): The number of documents in which a given term appear is calculated. This measure allows to determine how discriminative is a given term. Rare terms are more discriminative than common terms.
  - n-grams: For each weighting matrix, both TF and IDF are calculated for different sequences of words named n-grams. These text strings are the result of grouping together a sequence of words from a given text, previous preprocessing step. In this approach, we consider $n = 1, 2, 3$, i.e., word unigrams, bigrams and trigrams.
- Representative vector for each training data category-polarity:
  based on the weighting matrices generated with the training dataset, we proceed to create a representative vector for each category (Entity-Attribute) considering its associated polarity, for example, for the Entity ambience, the general attribute and its corresponding types of polarity (positive, negative, neutral and conflict), we obtain four different representative vectors:

$$\{AMBIENCE\#GENERAL, positive\}$$
$$\{AMBIENCE\#GENERAL, negative\}$$
$$\{AMBIENCE\#GENERAL, neutral\}$$
$$\{AMBIENCE\#GENERAL, conflict\}$$

- Detection by means of the cosine similarity measure:

We apply the cosine similarity measure, see Eq.(1), to determine the similarity between two weighting vectors, one of the training set and the other one from the test set. The target opinion will be assigned with the polarity according to the value obtained by the cosine measure (the highest one):

$$sim(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\boldsymbol{d}_j \cdot \boldsymbol{q}}{|\boldsymbol{d}_j| \times |\boldsymbol{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt[2]{\sum_{i=1}^{t} w_{ij}^2 \times \sum_{i=1}^{t} w_{iq}^2}}. \quad (1)$$

– Polarity evaluation. In order to determine the performance of the approach, we employ accuracy as the evaluation measure.

4 Obtained results

In this section we describe the results obtained with the proposed approach.

4.1 Dataset

In the experiments carried out, we use the training and test datasets provided by SemEval 2016, task 5, subtask 2. Test dataset includes the gold standard evaluations, so that it is possible to measure the quality of the approach proposed. User opinions are given for two domains: Restaurants (written in English and Spanish), and Laptops (written only in English). In Table 1 we show the number of texts (opinions) provided by SemEval 2016.

It is very important to mention that opinions in both, training and test datasets, may be assigned with more than one category and polarity. In Table 2, it is shown the number of tuples that opinions may have associated for each domain, i.e., the different categories and polarities by opinion.

Table 1. Number of texts given for each domain in subtask 2 of task 5.

Domain	TRAINING	TEST	GOLD
Restaurant (Spanish)	627	268	268
Restaurant (English)	335	90	90
Laptops (English)	395	80	80

Table 2. Number of tuples by domain.

Domain	TRAINING	TEST	GOLD
Restaurant (Spanish)	2,121	881	881
Restaurant (English)	1,435	404	404
Laptops (English)	2,082	545	545

The domain Restaurants-Spanish presents 12 categories with four possible polarities for each one. In Table 3 it is shown the corresponding information for each type of category and polarity in the domain of Restaurants (Spanish and English), see also Table 4.

Table 3. Distribution of polarity and category (tuples) for the domain Restaurants-Spanish.

Category	Positive		Negative		Neutral		Conflict	
	Gold	Train	Gold	Train	Gold	Train	Gold	Train
AMBIENCE#GENERAL	61	150	27	50	3	10	5	11
FOOD#QUALITY	148	383	17	51	11	17	7	10
FOOD#STYLE_OPTIONS	30	79	14	47	4	5	2	3
FOOD#PRICES	23	61	13	47	1	5	1	0
RESTAURANT#GENERAL	175	436	45	108	19	36	15	23
RESTAURANT#PRICES	17	50	19	44	3	13	0	1
RESTAURANT#MISCELLANEOUS	6	8	5	4	1	1	0	0
SERVICE#GENERAL	123	301	30	64	5	4	4	10
DRINKS#PRICES	3	4	5	6	0	0	0	0
DRINKS#QUALITY	9	20	0	9	0	0	0	0
DRINKS_STYLE#OPTIONS	5	11	5	8	1	0	0	0
LOCATION#GENERAL	18	13	0	2	0	0	0	0
TOTAL	618	1516	180	440	48	91	34	58

4.2 Experimental results

Taking into account the algorithm aforementioned, tuples are first evaluated by category and thereafter by polarity. In Table 5 it is presented the domain, the total of tuples per domain, the amount of samples classified by employing the TF text representation with unigrams (1-gram), bigrams (2-grams) y trigrams (3-grams), and the same when using TF-IDF. In Table 6, the results are reported with average accuracy for each domain.

The obtained results allow to determine that TF reports accuracies greater than 50% for domain Laptops, whereas TF-IDF obtains acceptable results when word unigrams are used.

5 Conclusions

In this paper it is presented an algorithm for automatic classification for identification of polarity and category for a given set of tuples provided by task 5 of SemEval 2016, in particular, by subtask 2. The domains considered for the tests are Restaurants (Spanish), Restaurants (English), and Laptops (English). According to the obtained results, the proposed algorithm obtained a performance

Table 4. Distribution of polarity and category (tuples) for the domain Restaurants-English.

Category	Positive		Negative		Neutral		Conflict	
	Gold	Train	Gold	Train	Gold	Train	Gold	Train
AMBIENCE#GENERAL	34	130	1	21	2	10	1	4
FOOD#QUALITY	69	235	9	47	4	13	4	19
FOOD#STYLE_OPTIONS	15	62	7	28	6	3	4	2
FOOD#PRICES	6	33	9	34	3	0	0	2
RESTAURANT#GENERAL	68	249	20	76	1	4	1	6
RESTAURANT#PRICES	6	33	9	25	1	5	0	0
RESTAURANT#MISCELLANEOUS	13	38	8	22	3	6	0	2
SERVICE#GENERAL	43	140	19	61	1	6	1	6
DRINKS#PRICES	0	14	1	6	0	0	0	0
DRINKS#QUALITY	15	33	0	4	0	2	0	0
DRINKS_STYLE#OPTIONS	10	27	1	2	0	0	0	0
LOCATION#GENERAL	7	18	0	1	2	6	0	0
TOTAL	286	1012	84	301	23	55	11	41

Table 5. Results of opinions classified correctly according to polarity by domain.

Domain	Test	TF			TF-IDF		
		1-*gram*	2-*grams*	3-*grams*	1-*gram*	2-*grams*	3-*grams*
Restaurant (Spanish)	881	589	506	459	535	570	466
Restaurant (English)	404	288	248	249	268	248	248
Laptops (English)	545	290	276	295	324	302	296

of 50% of accuracy. Unigrams is the text representation schema that obtained the best results for the Restaurants-English domain with an accuracy of 71%. TF obtained the best results with unigrams for the domain Restaurants-Spanish. Finally, unigrams with TF-IDF obtained the best results for Laptops-English with an accuracy of 59%.

Important is to mention that we have not employed any additional resource, such as dictionaries or lexicons for this classification process. We have only employed information provided by SemEval. As future work we plan to employ other linguistic resources as SentiWordnet as well as other text features for improving the performance of the approach.

Table 6. Accuracy results by domain.

Domain	TF			TF-IDF		
	1-*gram*	2-*grams*	3-*grams*	1-*gram*	2-*grams*	3-*grams*
Restaurant (Spanish)	**66.85**	57.43	52.09	60.72	64.69	52.89
Restaurant (English)	**71.28**	61.38	61.63	66.33	61.38	61.38
Laptops (English)	53.21	50.64	54.12	**59.44**	55.49	54.31

Acknowledgements. This research work has been partially supported by PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854, by VIEP-BUAP project 00570 and CONACyT project 257357.

References

1. Balaguer, E.V., Rosso, P., Locoro, A., Mascardi, V.: Análisis de opiniones con ontologıas. Polibits (41) pp. 29–36 (2010)
2. Brun, C., Perez, J., Roux, C.: Xrce at semeval-2016 task 5: Feedbacked ensemble modeling on syntactico-semantic knowledge for aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 282–286. Association for Computational Linguistics, San Diego, California (June 2016), TOBEFILLED-http://www.aclweb.org/anthology/W/W05/W05-0245
3. Deng, Z.H., Luo, K.H., Yu, H.L.: A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications 41(7), 3506–3513 (2014)
4. Hercig, T., Brychcín, T., Svoboda, L., Konkol, M.: Uwb at semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 354–361. Association for Computational Linguistics, San Diego, California (June 2016), TOBEFILLED-http://www.aclweb.org/anthology/W/W05/W05-0257
5. Martínez Cámara, E., Martín Valdivia, M.T., Perea Ortega, J.M., Ureña López, L.A.: Técnicas de clasificación de opiniones aplicadas a un corpus en español. Procesamiento del Lenguaje Natural 47(0), 163–170 (2011)
6. Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M.Á., Moreno, V., Fraga, A., Sánchez-Cervantes, J.L.: Feature-based opinion mining through ontologies. Expert Systems with Applications 41(13), 5995–6008 (2014)
7. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 19–30. Association for Computational Linguistics, San Diego, California (June 2016), http://www.aclweb.org/anthology/S16-1002
8. Saias, J.: Sentiue: Target and aspect based sentiment analysis in semeval-2015 task 12. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 767–771. Association for Computational Linguistics, Denver, Colorado (June 2015), http://www.aclweb.org/anthology/S15-2130