TASS 2015, septiembre 2015, pp 87-92 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

Comparing Supervised Learning Methods for Classifying Spanish Tweets

Comparación de Métodos de Aprendizaje Supervisado para la Clasificación de Tweets en Español

Jorge Valverde, Javier Tejada and Ernesto Cuadros
Universidad Católica San Pablo
Quinta Vivanco S/N , Urb. Campiña Paisajista, Arequipa - Perú
{andoni.valverde, jtejadac, ecuadros}@ucsp.edu.pe

Resumen: El presente paper presenta un conjunto de experimentos para abordar la tarea de clasificación global de polaridad de tweets en español del TASS 2015. En este trabajo se hace una comparación entre los principales algoritmos de clasificación supervisados para el Análisis de Sentimientos: Support Vector Machines, Naive Bayes, Entropía Máxima y Árboles de Decisión. Se propone también mejorar el rendimiento de estos clasificadores utilizando una técnica de reducción de clases y luego un algoritmo de votación llamado Naive Voting. Los resultados muestran que nuestra propuesta supera los otros métodos de aprendizaje de máquina propuestos en este trabajo.
Palabras clave: Análisis de Sentimientos, Métodos Supervisados, Tweets Españoles

Abstract: This paper presents a set of experiments to address the global polarity classification task of Spanish Tweets of TASS 2015. In this work, we compare the main supervised classification algorithms for Sentiment Analysis: Support Vector Machines, Naive Bayes, Maximum Entropy and Decision Trees. We propose to improve the performance of these classifiers using a class reduction technique and then a voting algorithm called Naive Voting. Results show that our proposal outperforms the other machine learning methods proposed in this work.
Keywords: Sentiment Analysis, Supervised Methods, Spanish Tweets

1 Introduction

Sentiment analysis is the computational study of opinions about entities, events, people, etc. Opinions are important because they often are taken into account in decision process. Currently, people use different social networks to express their experiences with products or commercial services. Twitter is one of the biggest repositories of opinions and it is also used as a communication channel between companies and customers. The data generated in Twitter is important for companies, because -- with that information -- they could know what is been saying about their products, services and competitors. In recent years, several researches of NLP have developed different methods to address the sentiment analysis problem in Twitter. The vast majority of works aim to classify a comment, according to the polarity expressed, in three categories: positive, negative or neutral (Koppel and Schler, 2006). The supervised classification algorithms are the most used methods to classify comments or opinions.

In this paper, we present a comparison of some supervised learning methods which have achieved good results in other research works. Analyzing the errors of those methods, we propose to use a class reduction technique and a voting algorithm (which take into account the results of supervised classifiers) to improve the classification of opinions in Twitter.

The rest of the paper is organized as follows: Section 2 summarizes the main works in

sentiment analysis. Section 3 describes our proposal and in Section 4 we describe the results that we have gotten. Finally, in Section 5, the conclusions of this work are presented.

2 Related Work

There are two general approaches to classify comments or opinions in positive, negative or neutral: supervised and unsupervised algorithms. Supervised classification algorithms are used in problems which are known a priori the number of classes and representative members of each class. The unsupervised classification algorithms, unlike supervised classification, do not have a training set, and they use clustering algorithms to try to create clusters or groups (Mohri, Rostamizadeh and Talwalkar, 2012).

The sentiment classification task could be formulated as a supervised learning problem with three classes: positive, negative and neutral. The most used supervised techniques in sentiment analysis are Naive Bayes (NB), Support Vector Machines (SVM), Maximum Entropy, etc. In most cases, SVM have shown great improvement over Naive Bayes.

Cui, Mittal, and Datar (2006) affirm that SVM are more appropriate for sentiment classification than generative models, because they can better differentiate mixed feelings. However, when the training data is small, a Naive Bayes classifier could be more appropriate. One of the earliest researches on supervised algorithms which classify opinions is presented in (Pang, Lee, and Vaithyanathan, 2002). In that work, authors use three machine learning techniques to classify the sentiment in movies comments. They test several features to find the most optimal set of them. Unigrams, bigrams, adjectives and position of words are used as features in those techniques. Ye, Zhang, and Law (2009) used three supervised learning algorithms to classify comments: SVM, Naive Bayes and Decision Trees. They use the frequencies of words to represent a document.

Most researches are focused for the English language, since it is the predominant language on the Internet. There are less works of sentiment analysis in Spanish opinions; however, Spanish is playing an important role. For Spanish comments, Perea-Ortega and Balahur (2014) present several experiments to address the global polarity classification task of Spanish tweets. Those experiments have focused on different feature replacements. The replacements were mainly based on repeated punctuation marks, emoticons and sentiment words. The proposal of Hernandez and Li (2014) is based on semantic approaches with linguistic rules for classifying polarity texts in Spanish. Montejo-Raez, Garcia-Cumbreras and Diaz-Galiano (2014) use supervised learning with SVM over the sum of word vectors in a model generated from the Spanish Wikipedia. Jimenez et al., (2014) developed an unsupervised classification system which uses an opinion lexicon and syntactic heuristic to identify the scope of Spanish negation words. San Vicente and Saralegi (2014) implement a Support Vector Machine (SVM) algorithm. That system combines the information extracted from polarity lexicons with linguistic features. For Peruvian Spanish opinions, Lopez, Tejada and Thelwall (2012) use a specialized dictionary with vocabulary of that country for Facebook comments. Lopez, Tejada and Thelwall (2012) proposed one of the first researches that analyze Peruvian opinions. In that work, authors use a basic method based on lexical resources to classify comments from Facebook.

3 Proposed Approach

This paper has two major objectives: First, we make a comparison of some of the main algorithms of supervised classification for Sentiment Analysis: Support Vector Machines, Naive Bayes, Maximum Entropy and Decision Trees. The second goal is to use a class reduction technique and then a voting algorithm to improve the accuracy of final results. The architecture of our system can be seen in Figure 1.

3.1 Comparison of Methods

In this paper we compare some classification methods in order to determine the performance of these algorithms in a set of opinions written by Spanish users. For the experiments, we used the four supervised classifiers described previously. The comparison of methods has the Training and Classification Phase. These phases will be explained below.

3.1.1 Training

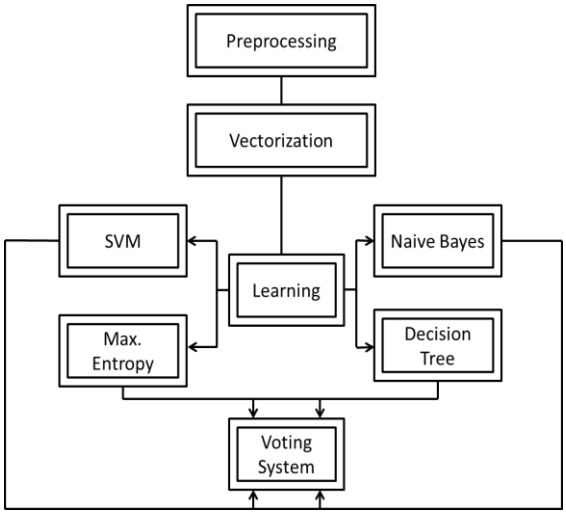

Figure 1: Proposed Approach

For each supervised classification methods used in this work, we identified three steps in the training phase: comment preprocessing, vectorization and learning.

Preprocessing: To make a correct comment preprocessing, we apply the following techniques:

- Elimination of symbols and special characters.
- Elimination of articles, adverbs, pronouns and prepositions (stopwords).
- Processing of hashtags.
- Correction of words with repeated letters.
- Filtration of words with ``@'' symbol as initial letter.
- Elimination of the characters ``RT''.
- URLs removal.
- Stemming of comments (opinions).

Vectorization: Each comment in the training data must be represented mathematically. There are different mathematical models to represent information. The most popular models are: boolean model, term frequency (TF), term frequency-inverse document frequency (TF-IDF) and Latent Semantic Analysis (LSA) (Codina, 2005). In this work, we decided to use the TF-IDF model to represent the comments of the corpus because it is more accurate and it has better results than the other models, (Salton and McGill, 1986). In Figure 2 it is shown an example of the corpus of tweets and its TF-IDF representation.

CORPUS OF TWEETS
Portada 'Público', viernes. Fabra al banquillo por 'orden' del Supremo; Wikileaks 'retrata' a 160 empresas espías.
Grande! RT @veronicacalderon "El periodista es alguien que quiere contar la realidad, pero no vive en ella"
Gonzalo Altozano tras la presentación de su libro 101 españoles y Dios. Divertido, emocionante y brillante
Mañana en Gaceta: TVE, la que pagamos tú y yo, culpa a una becaria de su falsa información sobre el cierre de @gaceta
Qué envidia "@mfcastineiras: Pedro mañana x la mañana me voy a Paris"
Más mañana en Gaceta. Amaiur depende de Uxue Barkos para crear grupo propio.

⇩

VECTOR SPACE REPRESENTATION OF TWEETS
(0,1),(1,1),(2,1),(3,1),(4,1),(5,1),(6,1),(7,1),(8,1),(9,1),(10,1),(11,2),(12,1),(13,1),(14,1)
(3,1),(9,1),(13,1),(15,1),(16,1),(17,1),(18,1),(19,1),(20,1)
(1,1),(2,1),(15,1),(21,1),(23,1),(24,1),(25,2),(26,2),(27,1),(28,1)
(16,1),(29,1),(30,1),(31,1),(32,1),(33,1)
(1,1),(11,1),(15,1),(29,1),(34,1),(35,1),(36,1),(37,1),(38,1)
(1,1),(2,1),(6,1),(11,2),(27,1),(29,1),(39,1),(40,1),(41,1),(42,1),(43,1),(44,1),(45,1)

⇩

TF-IDF WEIGHTING REPRESENTATION OF TWEETS
(0,031),(1,007),(2,0.12),(3,0.19),(4,0.31),(5,0.31),(6,0.19),(7,0.31),(8,0.31),(9,0.19, 10,0.31),(11,0.24),(12,0.31),(13,0.19),(14,0.31)
(3,0.26),(9,0.26),(13,0.26),(15,0.16),(16,0.26),(17,0.42),(18,0.42),(19,0.42),(20,0.42)
(1,0.07),(2,0.12),(15,0.12),(21,0.31),(22,0.31),(23,0.31),(24,0.31),(25,0.61),(26,0.31),(27,0.19),(28,0.31)
(16,0.29),(29,0.18),(30,0.47),(31,0.47),(32,0.47),(33,0.47)
(1,009),(11,0.16),(15,0.16),(29,0.16),(34,0.42),(35,0.42),(36,0.43),(37,0.43),(38,0.43)
(1,0.08),(2,0.13),(6,0.21),(27,0.21),(29,0.13),(39,0.34),(40,0.34),(41,0.34),(42,0.34),(43,0.34),(44,0.34),(45,0.34)

Figure 2: Vector Model Representation of Tweets

Learning: In this step, the classification algorithm receives as parameters the representative vectors of comments with their class labels. The class labels are: positive (P), negative (N), neutral (NEU) and none (NONE).

3.1.2 Classification

A classifier is a function that gives a discrete output, often denoted as class, to a particular input (Mohri, Rostamizadeh and Talwalkar, 2012). In this phase, the classifier receives a set of comments (the test data) and it evaluates this input to predict the corresponding class.

3.2 Our Proposal

In the first evaluation of the machine learning methods, the obtained accuracy results were slightly lower. For this reason, we propose to use two techniques to improve the results of classifiers. The first technique, called class reduction, removes one class label (NEU or NONE) with the aim of improving the margin of error of classifiers and reducing the number of classes to evaluate. The second technique, called naive voting, receives as input

parameters the optimized classifiers of the first technique. A more specific description of these techniques will be explained below.

3.2.1 Class Reduction

The basic idea of this technique was proposed in (Koppel and Schler, 2006). This technique is explained below.

Training and evaluation for three classes: We decided to train the classifiers considering three classes: Positive-Negative-Neutral and Positive-Negative-None. The classifiers were trained and tested in this way. The new results of classifiers using that simplification were better. Due to the improvement, we decided to join the partial results of these classifiers. With this union, we could classify the comments considering the four classes defined initially.

Union of partial results: We proposed to merge the partial results into single result. We established a set of rules to address the union of partial results of this class reduction technique, this rules are shown in Table 1.

Rule	Class Labels		Final Results
1	P	P	P
2	P	N	NEU
3	P	NONE	NONE
4	N	P	NEU
5	N	N	N
6	N	NONE	NONE
7	NEU	P	NEU
8	NEU	N	NEU
9	NEU	NONE	NONE

Table 1: Rules for Class Reduction Technique

3.2.2 Voting System

Our final technique presented in this paper was Voting System. We choose this method because all classifiers have a margin of error. Due to this margin of error, classifiers could classify incorrectly a comment. A voting system tries to reduce this margin of error. Voting systems are based on different classification methods. Many studies have used voting system to classify text. Kittler, Hatef and Matas (1998) and Kuncheva (2004) describe some of these methods. Rahman, Alam and Fairhurst (2002) show that in many cases the majority vote techniques are most efficient when classifiers are combined. Platie et al., (2009) and Tsutsumi, Shimada and Endo (2008) ensure that he following methods are the best voting systems for classification: Naive Voting, Weighted Voting, Maximun Choice Voting and F-Score/recall/precision Voting.

We proposed the Naive Voting technique, which has as input parameters the four classifiers proposed in this paper. Naive Voting is one of the simplest voting algorithms. In this technique, the comment is classified according to the majority agreement, i.e., the class with most votes in each classifier will be the winning class. The rules we applied for Naive Voting are described in Table 2.

Rule	Class Labels				Voting
	P	N	NEU	NONE	
1	4	0	0	0	P
2	3	0/1			P
3	2	0/1			P
4	0	4	0	0	N
5	0/1	3	0/1		N
6	0/1	2	0/1		N
7	0	0	4	0	NEU
8	0/1	3	0/1		NEU
9	0/1	2	0/1		NEU
10	0	0	0	4	NONE
11	0/1			3	NONE
12	0/1			2	NONE
13	2	2	0	0	P/N
14	2	0	2	0	NEU
15	2	0	0	2	NONE
16	0	2	2	0	NEU
17	0	2	0	2	NONE
18	0	0	2	2	NONE
19	2	0	2	0	NEU
20	1	1	1	1	NEU

Table 2: Naive Voting Rules

Each row of the Table 2 shows the votes obtained by each of the polarities (P-N-NEU-NONE) according to the output of the proposed classifiers. Due to we have 4 classifiers, the largest vote is 4 and the minimum is 0. Then, the class with the highest vote will be the winning class. In the event of a tie, a set of rules were established to determine the winning class. For example, in the case of a draw at 2 between positive and negative classes, a lottery was established to determine the winning. In other cases of a tie, it was chosen the NEU class or NONE class as the winner.

4 Experimental Results

4.1 Training and Test Data

We used the corpora provided by the organization of TASS 2015. For our purposes, we used the General Corpus and the Balanced General Corpus. The first one is composed of training and test set which contains 7219 and 60798 tweets, respectively. The Balanced General Corpus is a test subset which contains 1000 tweets only for test. A complete description of these corpora is explained in (Villena Román et al., 2015).

4.2 Evaluation of Classifiers

We performed a series of tests to address the Task 1 of TASS 2015, focusing on finding the global polarity of the Tweets corpora for 4 class labels (P-N-NEU-NONE). A general description of the "RUNs" that we have made for TASS 2015 are described in Table 3.

Tech.	Run-Id 60798	Run-Id 1000	Description
SVM	UCSP-RUN-2	UCSP-RUN-2	Support Vector Machine
NB	TestNB60000	UCSP-RUN-2-NB	Naive Bayes
ME	UCP-RUN-2-ME	TestME1000	Max. Entropy
DT	TestDT60000	TestDT10000	Decision Tree
SVM II	UCSP-RUN-1	UCSP-RUN-1	SVM + Class Reduction
NB II	UCSP-RUN-1-NB	UCSP-RUN-1-NB	NB + Class Reduction
ME II	UCSP-RUN-1-ME	UCSP-RUN-1-ME	ME + Class Reduction
DT II	UCSP-RUN-1-DT	UCSP-RUN-1-DR	DT + Class Reduction
Voting	UCSP-RUN-3	UCSP-RUN-3	Naive Voting

Table 3: Proposed Techniques

The results we have gotten for the evaluation of our proposal are shown in Table 4 (Evaluation of full test corpus) and Table 5 (Evaluation of 1k test corpus). It can be seen that class reduction techniques and our voting algorithm improve the accuracy of the original supervised classification algorithms.

Class reduction techniques improve results because they allow the classifier having to decide between fewer options and then the classifier could reduce its margin of error.

The voting algorithm gives good results because it takes into account the decisions of all the classifiers. This algorithm tries to reach a single decision that might be the best. A voting algorithm is like a consensus between all classifiers. But it is important to take into account that any voting algorithm is good as long as the majority of voters (classifiers) are good, otherwise, the voting algorithm will not have the expected results.

Approaches	Methods	Accuracy
Comparative	SVM	0.594
	NB	0.560
	ME	0.479
	DT	0.494
Proposal	SVM II	0.602
	NB II	0.560
	ME II	0.600
	DT II	0.536
	Voting	0.613

Table 4: Results of evaluating the Full Test Corpus

Approaches	Methods	Accuracy
Comparative	SVM	0.586
	NB	0.559
	ME	0.618
	DT	0.459
Proposal	SVM II	0.582
	NB II	0.636
	ME II	0.626
	DT II	0.495
	Voting	0.626

Table 5: Results of evaluating the 1k-Test Corpus

5 Conclusion

One of the main goals of this paper was to evaluate some supervised classification algorithms in the task of sentiment analysis. The results of evaluating the classifiers in initials experiments were not satisfactory. Using an optimization stage (class reduction and voting systems), accuracy improved

slightly compared to the original techniques. It could be shown that adequate voting algorithms improve the accuracy of classifiers. For proper operation of a voting system it is required to have multiple classifiers with a relatively high rate of efficiency. If a classifier fails, the other could give the correct prediction. But if most of classifiers give low results, then the voting system does not ensure a correct performance.

Acknowledgments

The research leading to the results has been founded by Programa Nacional de Innovación para la Competitividad y Productividad (Innóvate Perú)

References

Codina , L., 2005. Teoría de la recuperación de información: modelos fundamentales y aplicaciones a la gestión documental. *Revista internacional científica y profesional.*

Cui, H., V. Mittal and M. Datar. 2006. Comparative experiments on sentiment classification for online product reviews. *Proceedings of the 21st national conference on Artificial intelligence, Boston, Massachusetts.*

Hernandez, R. and Xiaoou Li. 2014. Sentiment analysis of texts in spanish based on semantic approaches with linguistic rules. *Proceedings of the TASS workshop at SEPLN 2014.*

Jimenez, S., E. Martinez, M. Valdivia and L. Lopez. 2014. SINAI-ESMA: An unsupervised approach for Sentiment Analysis in Twitter. *Proceedings of the TASS workshop at SEPLN 2014.*

Kittler, J., M. Hatef and J.Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Koppel, M. and J. Schler. 2006. The Importance of Neutral Examples for Learning Sentiment. *Dept. of Computer Science, Ramat Gan, Israel.*

Kuncheva, A. L.I. 2004. Combining Pattern Classifiers: Methods and Algorithms. *Jhon Wiley and Sons.*

Lopez, R., J. Tejada and M. Thelwall. 2012. Spanish Sentistrength as a Tool for Opinion Mining Peruvian Facebook and Twitter. *Artificial Intelligence Driven Solutions to Business and Engineering Problems.*

Mohri, M., A. Rostamizadeh and A. Talwalkar. 2012. Foundations of Machine Learning. *The MIT Press.*

Montejo-Raez, A., M.A. Garcia-Cumbreras and M.C. Diaz-Galiano. 2014. SINAI Word2Vec participation in TASS 2014. *Proceedings of the TASS workshop at SEPLN 2014.*

Pang, B., L. Lee and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10.*

Perea-Ortega, J. and A. Balahur. 2014. Experiments on feature replacements for polarity classification of Spanish tweets. *Proceedings of the TASS workshop at SEPLN 2014.*

Platie, M., M. Rouche, G. Dray and P. Poncelet. 2009. Is a voting approach accurate for opinion mining? *Centre de Recherche LGI2P.*

Rahman, A., H. Alam and M. Fairhurst. 2002. Multiple Classifier Combination for Character Recognition: Revisiting the Majority Voting System and Its Variation.

Salton G. and McGill M. 1986. Introduction to modern information retrieval.

San Vicente, I. and X. Saralegi. 2014. Looking for Features for Supervised Tweet Polarity Classification. *Proceedings of the TASS workshop at SEPLN 2014.*

Tsutsumi, K., K. Shimada and T. Endo. 2008. Movie Review Classification Based on a Multiple Classifier. *Department of Artificial Intelligence.*

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. 2015. Overview of TASS 2015.

Ye, Q., Z. Zhang and R. Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36:6527-6535*