Text Classification for Sentiment Analysis

Jose Mª Quinteiro-Gonzalez, Pablo Hernandez-Morera, Aarón López-Rodríguez
IUMA, Institute for Applied Microelectronics
University of Las Palmas de Gran Canaria, Spain
[jquinteiro, pablo, alopez]@iuma.ulpgc.es

Abstract—**The external opinions have always been a factor when taking a decision. Before the boom of blogs and social networks, we used to ask friends for their opinion, read magazines or ask the shop assistant before buying a car. Now we have more options. Blogs, opinion pages and social networks have increased the number of query options, generating a large amount of information, generally unstructured. And that is where the sentiment analysis and opinion mining.**

Sentiment analysis is to obtain the underlying polarity of a free text..Two approaches have been selected. Firstly let us consider the simplest approach and the most common in the literature field, the bags of words; and secondly the feature extraction .

Keywords-component; sentiment; opinion mining; bag of word; weka, sentiwordnet, tree-tagger

I. INTRODUCTION

Sentiment analysis is to obtain the underlying polarity of a free text. We understand as underlying polarity how positive or negative text can be. This process is also known as opinion mining and can be considered a branch of data mining by the use of its tools and techniques.

This Master Final Project considers the polarity of a text in order to classify it as positive or negative. In subsequent work, this classification can evolve into more complex concepts establishing ranges in the classification system. An example of such systems may be in stars, between one or five stars, or the establishment of polarity rates.

In order to create models and to compare them, two approaches have been selected. Firstly let us consider the simplest approach and the most common in the literature field, the bag of words, and secondly the feature extraction.

II. OUR APPROACH

A. Feature Extraction

For feature extraction TreeTagger is applied for tagging words,. For each document the number of verbs, nouns, adjective and adverb is determined and normalized by the number of token within the document under consideration.

Once we have the part of speech polarity is determined using SentiWordNet[1]. SentiWordNet has multiple entries for a word, even in the same part of speech. We decide to apply arithmetic mean. This approach extract feautres from text

- Frequency of positive and negative words (two attributes),
- Frequency of nouns, verbs, and adjectives (three attributes),
- Number of sentences, question marks and exclamation marks (three attributes)

We added new attributes:

- Frequency of adverbs (one attributes)
- Average polarity score triples for adverbs (three attributes)
- Number of negative symbols (one attributes)

Then we get 21 attributes.

B. Bag of Word

In this approach a text is represented as an unordered collection of words, disregarding grammar and even word order.

We will use n-gram with size from one to four, this will be use to see how size affects results. Therefore we will use stemmer and stopwords to improve frequency and to decrease words that only do noise. This consideration let us three degrees of freedom.

C. Algorithms

We use four machine leaning algorithm:

- *Naïve Bayes* [2] assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature
- *Bayes Multinomial* is a variation of naive bayes that considers not the frequency of articles of a class but the frequencies of the words in a class.
- . *J48* C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan.
- *Simple Logistic* , la regresión logística[3] se utiliza cuando es objetivo es describir la relación entre una clase categórica y un conjunto de variables asociadas, que pueden ser categorías o cuantitativas. En este caso son todas cuantitativas.

D. *Amazon Dataset*

The product review data set consists of Amazon product reviews that have already been used by Blitzer et al. [4]. Reviews are available for four different product types: books, DVD, electronics and kitchen appliances. Each review consists of a rating (0-5 stars), a reviewer name and location, a product name, a review title and date, and the review text. Reviews with ratings larger than 3 were labeled positive, those with rating smaller than 3 were labeled negative, and the rest was discarded because their polarity was ambigu- ous. For each domain around 1000 negative and 1000 positive labeled examples are available.

III. Results

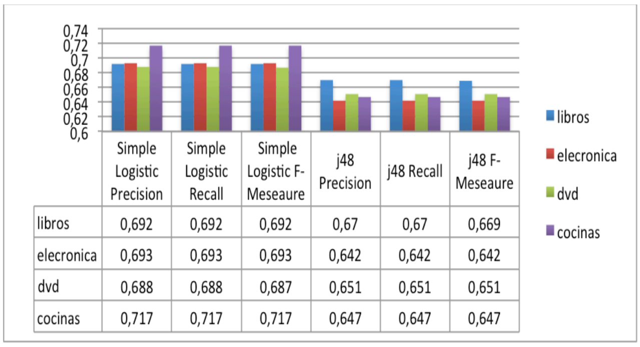
Although all algorithms tested in both approaches only show the data of the two best algorithms, which are also those who a priori are better suited to each approach.

First approach is a very poor approximation Fig. 1, because the results are far from the barrier of 80%, which is usually set as a goal. We can also observe that the logistic regression in all cases better than the J48 decision tree.

Second approach tests were conducted for each of the domains, kitchen, electronics, DVDs and books, although the graphics include all domains, use the same explanation as the results are very similar.

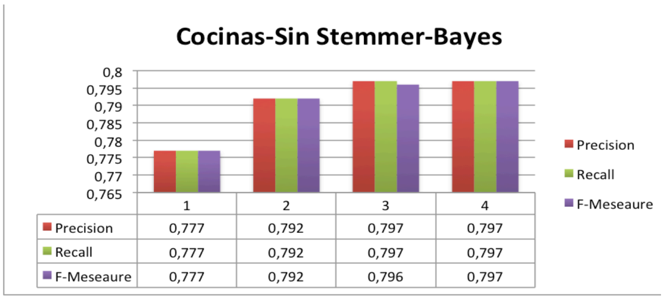
Fig 2 we see as in the case of not using stemmer and stopwords, the N-gram increase accuracy and far-f but from the use of N = 2 to N = 4 the improvement is minimal.

Fig 3 using the same parameters as above but the other algorithm under study we see something that is standard for all data N = 2 is the best option or at least as good as the use of N larger. We can also observe that for all cases Multinomial Bayes algorithm is better than Naive Bayes.



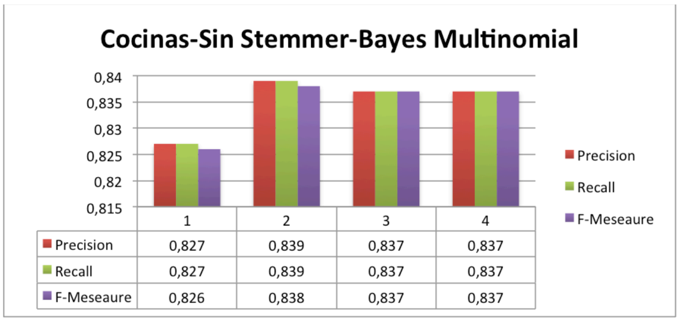
	Simple Logistic Precision	Simple Logistic Recall	Simple Logistic F-Meseaure	j48 Precision	j48 Recall	j48 F-Meseaure
libros	0,692	0,692	0,692	0,67	0,67	0,669
elecronica	0,693	0,693	0,693	0,642	0,642	0,642
dvd	0,688	0,688	0,687	0,651	0,651	0,651
cocinas	0,717	0,717	0,717	0,647	0,647	0,647

Figure 1. Feature Extraction



	1	2	3	4
Precision	0,777	0,792	0,797	0,797
Recall	0,777	0,792	0,797	0,797
F-Meseaure	0,777	0,792	0,796	0,797

Figure 2. Bag of Word Bayes Multinomial



	1	2	3	4
Precision	0,827	0,839	0,837	0,837
Recall	0,827	0,839	0,837	0,837
F-Meseaure	0,826	0,838	0,837	0,837

Figure 3. Bag of Word: Naïve Bayes

IV. Conclusion

This Master Final Project has met the goal of being an initiation into a new way of research. We have explored two ways of approach neglecting other equally valid.

There is evidence that a large number of N-gram classifier does not improve, and that the classical techniques of data classification, stemmer and stopwords If the best.

Has created an architecture that can be reused in the future, and common processes including the sentiment, that will save us many hours of work in subsequent investigations.

It has also allowed us to include different APIs under a single interface, allowing us to include new tools and compare them quickly and easily.

Finally, we have created a prototype classifier, which has improved but results in a human range as seen in the results section.

The following steps we must take is to improve the architecture, including tools for automatic selection of attributes and filtering the relevant N-grams. More Tools and Freeling parsing [6] and new APIs for input of data collected from rss and social networks like twitter could be added.

References

[1] Andrea Esuli and Fabrizio Sebastiani. SentiWordNet : A publicly available lexical resource for opinión mining. In *Proceedings of Language Resources and Evaluation (LREC)*, 2006.LANGLEY, P. IBA, W. y THOMPSON, K. An analysis of Bayesian classifiers. En: AAAI-92, (1992).

[2] Landwehr, N., Hall, M., & Frank, E. (2003). Logistic model trees. In Proc 14th European Conference on Machine Learning (pp. 241–252). Springer-Verlag

[3] M. D. John Blitzer and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for senti- ment classification. Association of Computational Linguis- tics (ACL), 2007

[4] FreeLing 2.2 (2011, Junio) http://nlp.lsi.upc.edu/freeling/