Sentiment of Financial News: A Natural Language Processing Approach

Leandro Alvim[#1], Paula Vilela[#2], Eduardo Motta [#3], Ruy Luiz Milidiú [#4]

[#]*Pontifícia Universidade Católica do Rio de Janeiro*
Brazil
[1]leandrouff@gmail.com
[2]pvilela@inf.puc-rio.br
[3]emotta@inf.puc-rio.br
[4]milidiu@inf.puc-rio.br

Abstract— **A huge amount of information is available online, in particular regarding financial news. Research in financial domain has shown that informational and sentiment aspects of stock news has a profound impact on market variables such as volume trades, volatility, stock prices and firm earnings. Here, we investigate the use of natural language pre-processing techniques in a way to improve the sentiment classification accuracy of a classical bag of words approach. We use the linguistic features provided by part-of-speech tagging, text chunking and negation. Entropy Guided Transformation Learning algorithm is applied to obtain the required features. For sentiment classification, Support Vector Machines and Naïve Bayes algorithms are used. Moreover, a Portuguese financial news annotated corpus is presented. It is composed by a collection of one thousand and five hundred newspaper reports about the Petrobras energy company. Our results show significant improvement of sentiment classification using Support Vector Machines when compared against the Naives Bayes strategy. The natural language techniques slightly improve the sentiment classification accuracy of the proposed models.**

I. Introduction

Recently, the interest in sentiment analysis has increased, since the scientific challenges became clear and the scope of new applications enabled by the processing of subjective language were noticed.

The task of classifying the sentiment of a text presents a large number of challenges. Frequently, the discourse is subtle, so, for instance, a negative review may not present a lot of negative words and vice-versa for a positive text.

For example, the sentence *"If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut."* doesn't include negative words but is a negative review of a perfume.

Another example of subtle discourse is *"This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up"*. This example presents many positive words but the last sentence inverts the whole concept.

The bag-of-words approach has been effectively applied to this problem [1,2]. Here, we propose a Natural Language preprocessing of the text that provides POS tagging, chunking and negation bigrams to enhance the bag-of-words approach.

II. Natural Language Approach

In order to experiment with sentiment classification performance, the following natural language features are considered.

A. POS Tagging

The Part-of-speech of each word is used to compose SVM [4] features. For the first experiments we count each POS class and considered this count as SVM feature for each document. On a second experiment we do a bigram of each word and its POS class.

B. Text chunking

Another linguistic feature that we use is the chunk type. For the first experiments we count each chunk type and provide as a SVM feature for each document. On a second experiment we do a bigram with each word and it chunk type.

C. Negation

Sentiment polarity classification can possibly benefit from negation structure preprocessing. Bigrams are constructed by including the word NOT and the next word. For example, in the sentence *“I don’t like this movie”*, the bigram NOT_like is constructed resulting in *“I do NOT_like this movie”*.

III. Financial News Corpus

We introduce PETRONEWS, a sentiment annotated corpus of Portuguese news about Petrobras [1]. PETRONEWS is a collection of one thousand and five hundred newspaper reports about that Brazilian company, automatically collected from the *Gazeta Mercantil* and *Valor Econômico* websites. All these news reports have been annotated as having either a positive or negative sentiment.

PETRONEWS has features of two types for each financial news: bag-of-words and Natural Language processing features. For building the bag-of-words features, we use a dictionary that contains all the corpus words. Each document is represented by a sparse binary array indicating dictionary word presence. For the Natural Language features processing, we use the Entropy Guided Transformation Learning algorithm [5].

IV. Results

As a dataset baseline for our experiments, we work with the movie reviews dataset from Epinions.com, used by Pang and Lee [2]. This dataset contains 2,000 movie reviews, half positive and half negative. Other researchers have used this dataset, so we already have results to compare with ours. We report their result as the first and second model in Table I.

Both datasets have equal number of positive and negative texts, and for both we performed 5-fold and 10-fold cross validation. Punctuation is also included as a token in order to compare our results as directly as possible to previous ones.

For all experiments, SVM is used. This technique has been shown to be very effective for this task, providing better results than Naïve Bayes (NB) models. For PETRONEWS we use NB as a baseline technique to compare against the SVM results.

In the experiments with the movie review dataset, we obtain the same accuracy as Mullen and Collier [3]. We consider just the presence of a word. When bigram with the word and its POS tagging is considered, better results are achieved. On the other hand, when bigrams of the word and its chunk are used, the result worsened compared to just the presence feature. For all experiments, we considered only the words that appear at least 5 times in all examples. Results are listed in Table I. The first line shows the results of the work of Pang *et al* [2], where they propose this corpus. In the second line we show the results of Mullen and Collier [3] and Osgood *et al* [6], where they use a hybrid SVM which combine unigram-style feature-based SVM with those based on real-valued favorability measures, this was the best results yet published using this data. On the following lines we show our experiments’ results. The best one is the last line where we do a bigram with the word and it POS tagging and use as a feature for SVM.

TABLE I
ACCURACY RESULTS FOR THE MOVIE REVIEW DATASET

Model	**3 folds**	**5 folds**	**10 folds**
Pang *et al* 2002	82.90%	N/A	N/A
Hybrid SVM (PMI Osgood and Lemmas)	84.60%	85.00%	86.00%
SVM with Presence	84.45%	85.00%	86.00%

[1] Petrobras is a Brazilian multinational energy company.

SVM with Bigrams of Presence and Negation	N/A	83.50%	N/A
SVM with Count of POS tag and Chunk	N/A	85.00%	N/A
SVM with Bigrams of Presence and chunk	N/A	83.74%	N/A
SVM with Bigrams of Presence and POS tag	**84.80%**	**85.76%**	**86.09%**

For the experiments with Petrobras' corpus a Naïve Bayes classifier is used as baseline to compare the results with the SVM experiments. As shown in Table II, there is a considerable improvement over the NB when using SVM. For this corpus, when we use bigrams with POS tagging the results do not improve over just presence. As this dataset is smaller than the first one, words that appear at least 2 times in all examples are considered.

TABLE II
ACCURACY RESULTS FOR THE PETROBRAS' NEWS CORPUS

Model	**Accuracy (5 folds)**
Naïve Bayes	76.09%
SVM with Bigrams of Presence and POS tag	84.00%
SVM with Presence	**85.94%**

V. CONCLUSIONS

In this work, we investigate the use of part-of-speech tagging, text chunks and negation applying Natural Language preprocessing techniques in a way to improve the sentiment classification accuracy of a classical bag-of-words approach. We also introduce PETRONEWS, a Portuguese financial news sentiment annotated corpus.

Our results show significant improvement of sentiment classification when using Support Vector Machines as compared to a Naive Bayes strategy. The Natural Language approach slightly improves sentiment classification accuracy.

For the Natural Language Processing task it appears to remain considerable room to improvement. We plan to extend our approach by using Semantic Role Labeling to provide additional linguistic features.

ACKNOWLEDGMENT

This work was partially funded by CNPq and FAPERJ grants 557.128/2009-9 and E-26/170028/2008.

The first and third authors were supported by a CNPq doctoral fellowship.

REFERENCES

[1] Bo Pang and Lillian Lee. 2002. Opinion Mining and sentiment analysis

[2] Bo Pang, Lillian Lee and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques.

[3] Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources.

[4] Bernhard Boser, isabelle Gujon and Vladimir Vapnik. 1992. A Training algorithm for optimal margin classifiers

[5] Cícero N. dos Santos, Ruy L. Milidiú and Raúl P. Renteria.2008. Portuguese Part-of-Speech Tagging using Entropy Guided Transformation Learning

[6] Charles E. Osgood, George J. Succi, and Percy H.Tannenbaum. 1957. The Measurement of Meaning University of Illinois.