UNIVERSITÀ DI PISA

Text Analytics - Project Report

Michele Andreucci 628505

Carlo Paladino 537650

Giulia Calvo 544434

MSc in Data Science and Business Informatics

Anno accademico: 2021/2022

INTRODUCTION

"Sentiment Analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes and emotions from written language."[1]

SA is a special scenario for text Analytics Problems and it faces multidisciplinary approaches like NLP, Information Retrieval and ML to solve these problems: the aim is to capture how sentiments are expressed in natural language, knowing that there are many components in the language that can help to capture this, such as subjects, objects, attributes, verbs. With SA tools we try to recognize and model them into semantic Abstractions.

In this paper the aim is to analyze tweets from Kaggle[2] dataset composed by Financial news and the Sentiment associated with it, make our hypothesis and use ML and NLP models to test them, comment the results of the models that we found and apply the best model on "fresh data" to see how it performs.

3.2 EXPLORATORY DATA ANALYSIS

The dataset used in this project is composed by a main file called All data that contains 4846 rows and two columns, News and Sentiment: there are three sentiments "Neutral", "Negative" and "Positive" and in Figure 1 we can see the distribution of each of them.

Using different libraries, we found that:

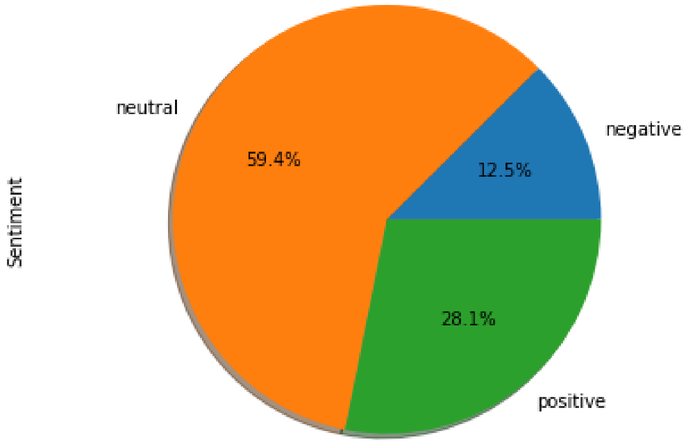

Figure 1: data distrubution by sentiment

- There are no repeated tweets and no missing values.
- All tweets are in English (even if some terms are in other languages)

[1] Bing Liu, "Sentiment Analysis and Opinion Mining" Morgan & Claypool Publishers, 2012.

[2] https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news

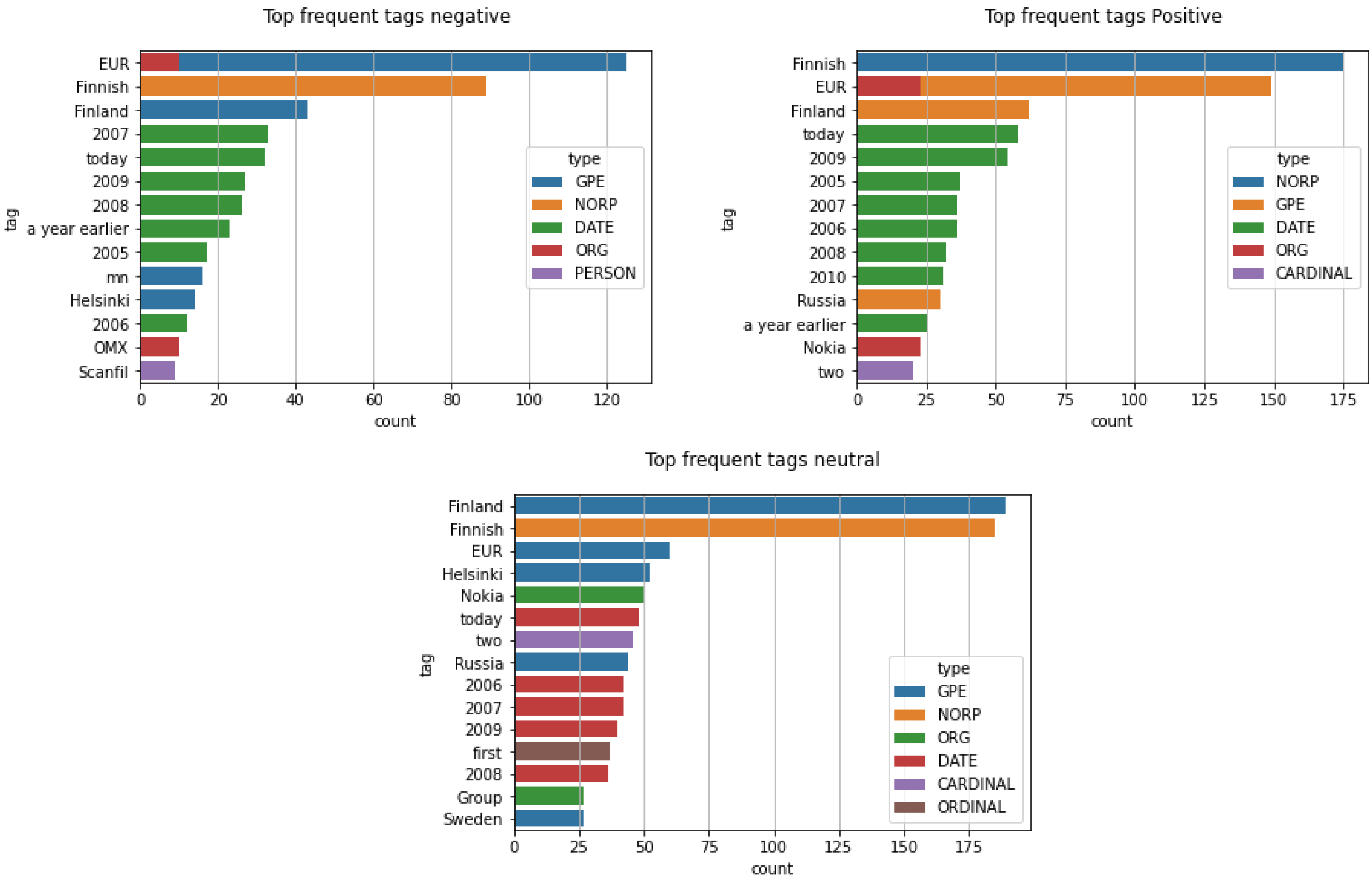

Table 1: Distribution of Tags for each Sentiment

Analyzing the table obtained using Spacy, we can see that for all the labels the type "Date" is predominant: this result was predictable because speaking in Financial terms, time is crucial to measure the performances of a company; this is supported also by the fact that other tags like Cardinal, Ordinal (crucial in the Financial analysis to monitor performances of companies) are also in the top frequent tags.

The other Tags are about Companies, agencies, institutions (ORG) and Geopolitical entity (GPE) and nationalities (NORP).

To demonstrate furthermore that "Date" type is predominant and strictly connected to performances of the companies, we also analyzed Unigrams and Bigrams most frequent words (Table 2): we decide to report the data only for positive label.

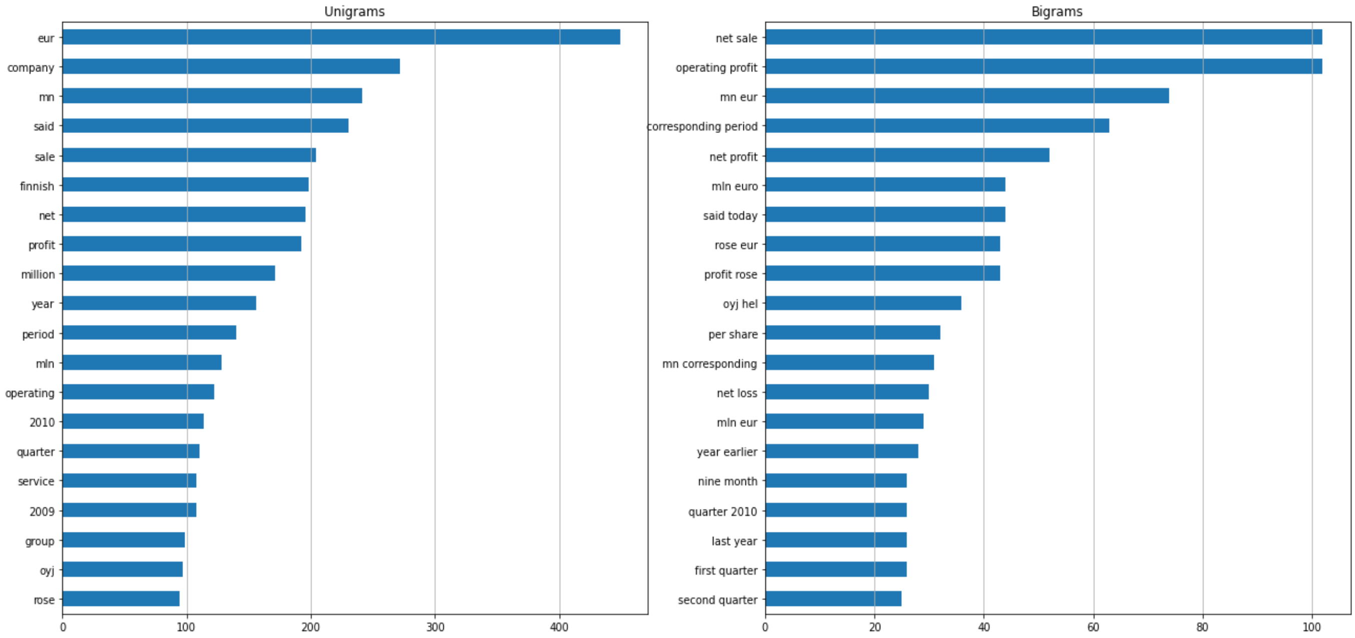

Table 2: Most frequent words for Positive News

We can see from the analysis of Bigrams that most of them are strictly connected with results from quarter performances (like first quarter, second quarter, etc.) or results from the financial statements (like net sale, operating profit, profit rose, net loss etc.).
What we found is in line with the dataset because it contains the sentiments for financial news headlines: financial investors are able to have an idea of the performances of the companies thanks to the bigrams found.

We also tried to find the 3 main topics inside the dataset and print for each of them 10 main unigram or bigram.

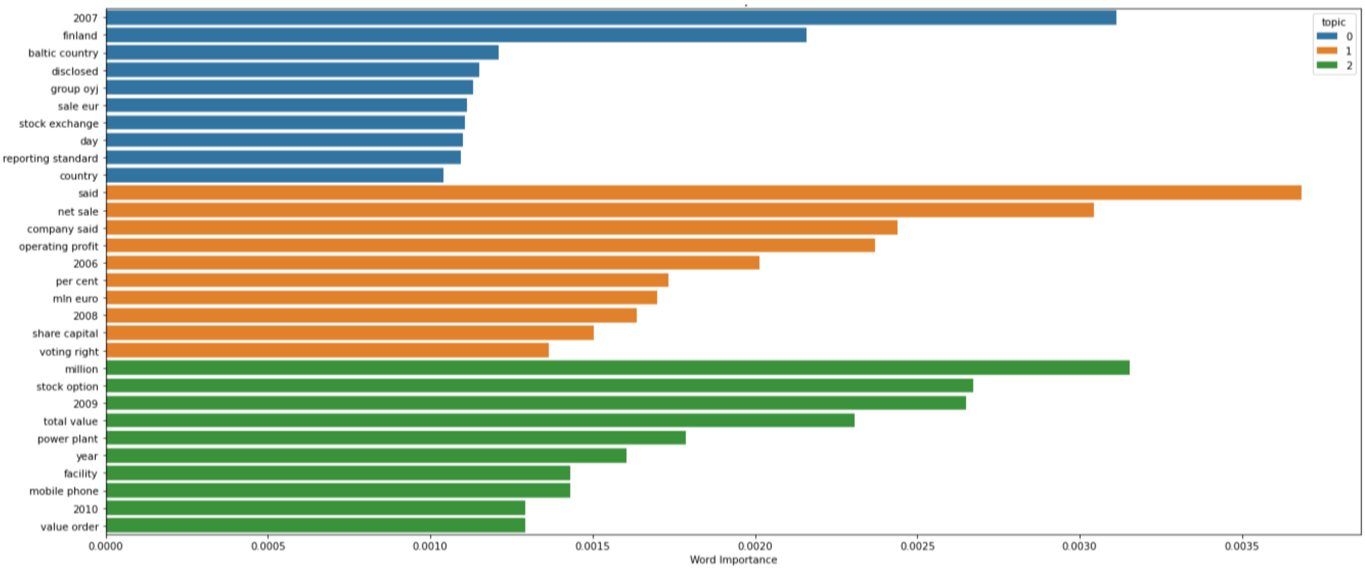

Table 3: 10 most unigrams and bigrams inside the main three Topics

3.2 DATA PREPARATION

Together with all the data, we received also other 4 files: each of them is a subset of the whole dataset but they differ in the level of agreement of the annotators. For example, the file “sentences_50agree” contains the sentences on which at least 50% of the annotators agreed on the sentiment label to be assigned: the percentage represents the level of agreements of the annotators.

For the process of data preparation (Table 3) of the subsets, we decide:

	Rows before Preprocessing	Rows after Preprocessing
Sentences_50agree	4846	3392
Sentences_66agree	4217	2763
Sentences_75agree	3453	1999
Sentences_Allagree	2264	810

- to take 30% of “sentences_50agree” as Test set
- to use all the other subsets as training sets
- to remove from the training sets the tweets contained in the test set

3.2 DATA MODELLING

We decide to explore the classification emphasizing the difference between Binary Classification (classifying only positive and negative records) and Multiclass Classification: what we want to see and to prove is that even if we know from theory that the Binary performs better than Multiclass classification, we can at least achieve good results with Multiclass.

We decide to explore different algorithms for the classification task starting from Logistic Regression, SVM and Naïve Bayes Classifier, then different Neural Networks ending with the BERT model.

Using Imbalance Learning we just made a comparison to see how the results with balanced dataset will be.

3.1 LOGISTIC REGRESSION, SVM AND NAÏVE BAYES CLASSIFIER

	Accuracy		*Precision*		*Recall*		*f-1 score*	
NaiveBayes								
	Binary	*Multiclass*	*Binary*	*Multiclass*	*Binary*	*Multiclass*	*Binary*	*Multiclass*
50Agree	0,79	0,72	0,75	0,55	0,77	0,64	0,76	0,63
66Agree	0,80	0,71	0,76	0,64	0,77	0,61	0,77	0,62
75Agree	0,80	0,71	0,76	0,64	0,74	0,59	0,75	0,60
All Agree	0,79	0,70	0,76	0,65	0,72	0,57	0,73	0,60
SVM								
	Binary	*Multiclass*	*Binary*	*Multiclass*	*Binary*	*Multiclass*	*Binary*	*Multiclass*
50Agree	0,84	0,79	0,83	0,84	0,79	0,65	0,8	0,7
66Agree	0,82	0,79	0,8	0,8	0,77	0,66	0,78	0,7
75Agree	0,83	0,77	0,82	0,77	0,76	0,63	0,78	0,67
All Agree	0,79	0,74	0,76	0,73	0,72	0,6	0,73	0,64
Logistic Regression								
	Binary	*Multiclass*	*Binary*	*Multiclass*	*Binary*	*Multiclass*	*Binary*	*Multiclass*
50Agree	0,81	0,75	0,84	0,8	0,71	0,58	0,73	0,62
66Agree	0,82	0,74	0,83	0,77	0,73	0,57	0,75	0,61
75Agree	0,81	0,73	0,83	0,76	0,71	0,55	0,73	0,59
All Agree	0,79	0,71	0,81	0,75	0,69	0,53	0,71	0,56

From this table we can easily affirm that the classifier with the highest Gap in terms of measures between binary and Multiclass Classification is the Naïve bayes Classifier probably since the classifier is based on Probabilities and the datasets are imbalanced.

In the Logistic Regression we can achieve closer results with respect to NB on Binary Classification and better ones on Multiclass: we assume this is due to the fact that it’s better than Naïve Bayes in dealing with probabilities.

The SVM instead did a good job for both Binary and Multiclass thanks to the library LinearSVC[3].

3.2 NEURAL NETWORKS

Arrived at this point we decide to use Neural Networks: we used first a simple neural network then Convolutional and Recurrent.

These are the Result for the Binary classification:

	ACCURACY		
Test set	*Simple Neural Networks*	*Convolutional Neural Network*	*Recurrent Neural Network*
50Agree	0,72	0,74	0,71
66Agree	0,72	0,78	0,71
75Agree	0,71	0,72	0,71
All Agree	0,71	0,72	0,71

As we can see, the CNN goes better than the other ones, but looking at the notebook it overfits our training sets; instead, the RNN achieves an average accuracy of 0,83 on training without overfitting.

We also tried Multiclass classification, but we didn’t achieve consistent results: we can affirm that for Neural networks the Binary classification gives better performances.

3.3 IMBALANCE LEARNING ON MULTICLASS PROBLEM

We also went through Imbalance Learning problem using Over or Under Sampling algorithms so that we could achieve better results on Multiclass Classification.

We then decided to use Logistic regressor on the dataset 50 Agree (the biggest one) to see how it goes on test set.

[3] https://scikit-learn.org/stable/modules/svm.html

Test set	*Accuracy*	*Precision*	*Recall*	*F1 - Score*
Over Sampling				
Random over sampler	0,76	0,76	0,72	0,76
SMOTE	0,77	0,68	0,66	0,67
Under Sampling				
Random under sampler	0,68	0,58	0,65	0,61
Near Miss 1	0,59	0,55	0,58	0,52
Near Miss 2	0,71	0,60	0,63	0,60
Near Miss 3	0,65	0,55	0,62	0,57

The results in this table were not expected: we expected at least on Under sampling better classifications, but Random oversampling had the best results, worse than the logistic regressor in terms of Precision but better in terms of Accuracy, Recall and F1-Score.

3.4 BERT

BERT (Bidirectional Encoder Representations from Transformers) uses the Transformer encoders as building blocks. The model is bidirectional in the sense that the attention model can peek at both left and right contexts.

Knowing this, we expected better results on it; this model takes a lot of time for training, so we decide to run it on the biggest dataset (50 Agree) doing Binary and Multiclass Classifications.

Test set	*Epochs*	*Accuracy*
Binary	2	0,90
Multiclass	4	0,80

We can affirm that the model works better than the others achieving a high accuracy with small number of epochs: this since is more likely to predict negative News even if this class represents only the 12,5% of the dataset.

We think that here the classification easily captures the context of the sentence because BERT can be finetuned with less resources on smaller datasets to optimize its performance on specific tasks and it's perfect for our dataset.

3.5 CRAWLER ON TWEETS

Using the API of Twitter we create a new crawler made of Financial Tweets: using "Operating Profit" as keyword (one of the most present bigrams inside the original dataset) we collected about 4500 tweets. After dropping the retweets, not useful for the prediction, we get 799 tweets to analyze. We decide to apply the Bert model on the dataset previously generated. In the end even if the accuracy of the prediction is satisfying, we notice that the model is not good in the prediction of the sentiment of the test set tweets.

CONCLUSIONS

After a really deep analysis of the different classification models, we demonstrate that our hypothesis is correct: the binary classification gives us better results that the multiclass, but multiclass gives acceptable results.