Tweet Sentiment Analysis Regarding the Brazilian Stock Market

Murilo C. Medeiros[1], **Vinicius R. P. Borges**[1]

[1]Universidade de Brasília - Departamento de Ciência da Computação
Brasília – DF – Brazil

silver472@me.com, viniciusrpb@unb.br

Abstract. *This paper describes a methodology for analyzing sentiments and for knowledge discovery in tweets regarding the Brazilian stock market. The proposed methodology starts by preprocessing and characterizing tweets to obtain an associated vector-space model. After that, a dimensionality reduction is employed by using Principal Component Analysis and t-Stochastic Neighbor Embedding. Sentiment analysis of stock market tweets is performed by considering the tasks of sentiment classification, topic modeling and clustering, along with a visual analysis process. Experiments results showed satisfactory performances in single and multi-label sentiment classification scenarios. The visual analysis process also revealed interesting relationships among topics and clusters.*

1. Introduction

Companies employ many people, create and develop many products and services, impacting people's lives around the world. The stock market plays an important role by representing the overall market performance of companies. Thus if a stock index composed by many companies is not performing well, then it is an indicator that the economy may not be fully using its productive capacity. This could mean that less people may be employed, resulting in people's personal lives being negatively affected by the performance of companies in the economy.

In the context of business and stock market, some investors often express their opinions about stocks on social networks and can make decisions based on the information presented in their personal profiles. One of the most popular social networks worldwide is Twitter[1], in which each user has a personal profile and can post short text messages (up to 280 characters), named *tweets*, to be presented to a list of friends.

The wide diversity of user profiles, their different opinions and the interesting aspects of tweets have been explored by the sentiment analysis field in several knowledge domains, such as education [Altrabsheh et al. 2014], healthcare [Clark et al. 2018], politics [Barberá and Rivero 2015] and recommendation systems [Li et al. 2016]. Specifically in Twitter, such users can send real-time information concerning facts and behaviors in the stock market community. For instance, if an important person, such as a CEO of a big company or a country's President, expresses his opinions about a particular stock, investors are likely to be influenced by taking further decisions, so that their actions will directly impact the stock price.

In literature, some researchers proposed strategies for analyzing sentiments of the stock market texts and opinions. However, few researches have focused on the Brazilian

[1]http://twitter.com/

stock market [L.Lima et al. 2016]. One of the reasons is due to the limited availability of social network data for this market compared to the stock market of other countries, particularly the United States. Some of these researches analyzed the influence of news from Brazilian newspapers on economy, but disregarded social network data [Galdi and Gonçalves 2018], especially those from Twitter.

In order to fill this gap, this paper proposes a methodology for analyzing Portuguese tweets concerning the main index in the Brazilian stock market. Our methodology was devised by considering a dataset of tweets related to the Brazilian stock market [Silva 2018], which were labeled according to several sentiments, such as joy, trust, fear, surprise, sadness, disgust, anger and anticipation. The proposed methodology employs some typical sentiment analysis tasks, such as sentiment classification, topic modeling and tweet clustering. Moreover, a visual analysis process for knowledge discovery on tweets is developed using two popular visualizations based on point placement strategies: Principal Component Analysis [Jolliffe 2011] and t-Stochastic Neighbor Embedding [Maaten and Hinton 2008].

This paper is organized as follows. Section 2 describes the related works that focused on sentiment analysis using data from social networks and other news sources. Section 3 explains the steps of the proposed methodology for sentiment analysis of stock market tweets. Section 4 discusses the experiments and the obtained results by the proposed methodology. Section 5 plans possibilities for future work.

2. Related Work

Several researches have explored social networks to analyze sentiments regarding the stock market. Generally, sentiment analysis can be performed according to machine learning, lexicon and hybrid approaches [Giachanou and Crestani 2016]. Our research focuses on machine learning approaches, in which predefined mathematical models learn the relevant patterns of different data categories from the a given labeled training set.

[Rao and Srivastava 2012] proposes to correlate sentiments contained in tweets with parameters of specific stocks or stock indexes. The article improves previous work in aspects such as tweet collection, sentiment classification, feature extraction and financial data collection. It introduces a new feature named "bulishness". The results presented improvements in relation to previous works and the methods are innovative and scalable. The article lacks on analyzing data in more details by considering other statistical and classification models.

[Bollen and Mao 2011] investigates whether a tweet sentiment in one or more dimensions is correlated with the close value of the Dow Jones index over time. The article proposes two approaches for sentiment analysis by verifying the linear correlation with the index. It applies a fuzzy neural network to verify the sentiment combinations and the non-linearity of the relation between the sentiments and the close value of the Dow index. The results indicated that some sentiments do not correlate with the index close value. Moreover, the authors showed the correlation between the sentiments, but the causal mechanisms connecting the sentiments to the closing values of the index are not indicated.

[Smailović et al. 2013] explored whether tweets are valid data sources for prediction of stock movement by employing support vector machines for sentiment classification according to the the neutral, positive and negative polarities. A causality analysis was conducted in order to show that tweet sentiment can be considered for stock movement

prediction. The experiment disregarding the neutral sentiment presented better results for stocks with many price variations. The experiment presenting the neutral sentiment achieved a better performance for stocks without many price variations, since the inclusion of the neutral sentiment enhanced the positive and negative polarities. The article could not unify a model for different kinds of stock tendencies.

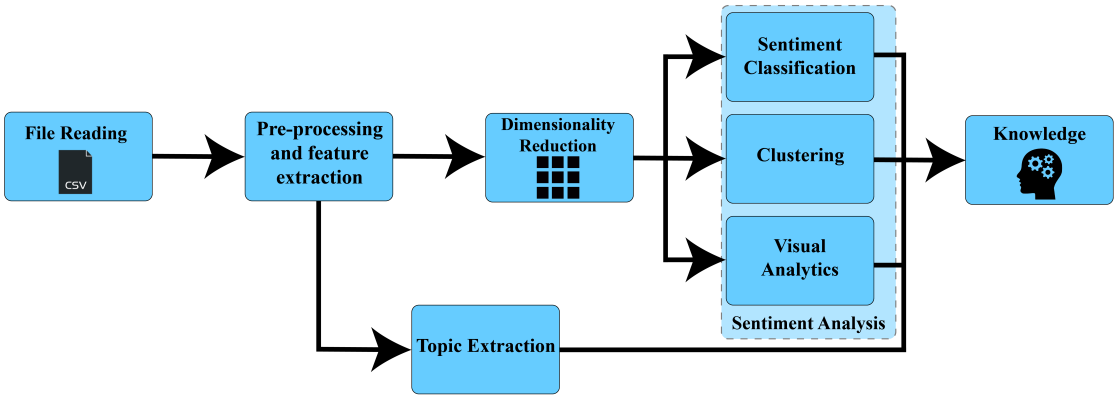
[L.Lima et al. 2016] presents a mechanism that uses natural language processing to determine the collective sentiment about stocks and extract patterns. The article collected tweets about Petrobras, the main Brazilian oil company, and applied many tools for sentiment analysis. The research reported that the collective humor about a stock does not reflect the real sentiment about this stock. Moreover, companies presenting a strong relation with politics cannot be affected by collective sentiments. However, the article considered only a single company in the analysis.

Few works were dedicated to analyze sentiments taking into account the Brazilian stock market. In order to fill this gap, we propose a methodology that receives as input a set of tweets related to the main Brazilian stock index (*Ibovespa*). As the strategy is to process and extract implicit knowledge from those tweets, feature extraction, classification and topic extraction techniques are employed to analyze and predict the multiple sentiments of Brazilian tweets.

3. Proposed Methodology

The flowchart summarizing the steps of the proposed methodology is depicted in Figure 3. The methodology receives the original set of tweets as input, which is appropriately preprocessed for further analysis. After that, the preprocessed tweets are converted to the bag-of-words representation using the TF-IDF approach. The TF-IDF vectors are used for topic modeling approaches, while a low dimensional representation of TF-IDF space is obtained by performing dimensionality reduction. Such representation is used as input to the clustering and classification techniques, allowing us to analyze the tweet sentiments. Finally, the performance of the classification tasks and the obtained implicit knowledge are reported and discussed.

Figure 1. The proposed methodology for tweet sentiment analysis.



3.1. Dataset

The dataset used in our research contains 4516 tweets about the Brazilian Stock Market and it is available in a public domain [Silva 2018]. All tweets were labeled according to

the Plutchik's Psychoevolutionary Theory of Basic Emotions [Plutchik 1980]: joy, trust, fear, surprise, sadness, disgust, anger and anticipation. If no prevailing sentiment was found in the tweet, it was labeled as neutral. The annotation process consisted by defining annotators to label each tweet in the dataset. Each annotator was able to select different labels for a given tweet and to label multiple tweets. Each tweet may contain more than one label, so an instance could be labeled for more than one sentiment, except for those which are considered opposite. Additional details concerning the annotation process can be obtained elsewhere [Silva 2018].

However, some tweets might appear as non-labeled, indicating that no agreement was achieved between the annotators. In this sense, only 1979 tweets had an agreement between the annotators. The number of instances for each sentiment is the following: 542 for Trust, 353 for Disgust, 332 for Joy, 236 for Sadness, 413 for Anticipation, 286 for Surprise, 157 for Anger, 110 for Fear and 779 for Neutral. It can be seen that the dataset is unbalanced, in which some sentiment categories may present more than the double of counts in relation to others.

3.2. Preprocessing and Feature Extraction

As tweets are prone to different variety of words, links and terms, a text preprocessing step is required to prepare them for further steps. First, the non-labeled tweets are removed in relation to a specific sentiment. Furthermore, tweets presenting irrelevant words regarding the stock market context are removed.

For each tweet, initially, all the words are converted to lowercase and are compared to a list of Portuguese stopwords [2]. Also, words containing two or less characters and URLs are not considered. Thus, a new string is constructed by removing stopwords and words with the characteristics mentioned above.

We performed preliminary experiments which reported that words such as "http" or "rt" were usually the main topics or belonged to clusters' centroids after performing topic extraction and clustering, respectively. The preprocessing allowed to produce relevant results that contained important words that expressed the main idea of the obtained topics and clusters. The feature extraction comprises of obtaining a structured representation of the preprocessed tweets by means of a vector-space model. In this case, each tweet is denoted by a feature vector in a multidimensional space, so that each dimension is related to a term in the dataset. A well-known approach is the bag-of-words model, in which for a given dataset constituted by n single terms of the dataset's vocabulary, each tweet t_i is represented by a vector $\vec{d_i} = (a_{i,1}, a_{i,2}, \ldots, a_{i,n})$, in a way that $a_{i,j}$ expresses the relation between the term j to the tweet t_i. Finally, the term frequency-inverse document frequency (TF-IDF) is computed for each vector, indicating the term's relevance regarding the total set of tweets.

The "cut-off" parameter, that is, the minimum document frequency considered in the TF-IDF vectorizer was 0.95, representing the proportion of documents. In this sense, terms presenting a document frequency lower than this threshold are ignored. Moreover, terms presenting more than two document frequencies are ignored.

3.3. Dimensionality reduction

TF-IDF vectors can present a high dimensionality since it depends on the number of terms in the entire set of tweets, whichcan affect the performance of machine learn-

[2]ISO Stopwords - `https://github.com/stopwords-iso/stopwords-pt`

ing algorithms [Zimek et al. 2012]. The dimensionality of the TF-IDF vectors were reduced using two well-known techniques in literature: Principal Component Analysis (PCA) [Jolliffe 2011] and t-distributed Stochastic Neighbor Embedding (t-SNE) [Maaten and Hinton 2008]. These approaches were originally formulated for dimensionality reduction, but in literature they have been successfully employed as visualization techniques. Therefore, PCA and t-SNE were set to produce 2D low dimensional spaces, so that we can also use them in the visual analysis process of stock market tweets.

PCA is a dimensionality reduction technique that performs a linear mapping of the data, defined in a high dimensional space, to a lower-dimensional space. First, data instances are centered by subtracting the mean values and the covariance matrix is computed. After that, such covariance matrix is eigendecomposed, thus obtaining the eigenvectors and eigenvalues. The greatest eigenvalues are identified in order to select the correspondent eigenvectors, defining the principal components. The transformed data is the result of the product between the principal components and the original data. Typically, the eigenvectors associated with the higher eigenvalues retain the most variability of data, so PCA attempts to maximize data's variance in fewer dimensions.

t-SNE is a visualization technique based on point placement that receives as input a high-dimensional dataset and generates a graphical representation from a matrix pairwise similarities. This nonlinear technique employs a Student-t distribution to compute the similarity of Euclidean distances between data points in the low dimensional space. t-SNE presents interesting properties, since it preserves the similarity relations among neighboring data instances, as well as it reveals clusters with different patterns and the relationships among them.

3.4. Topic modeling

The goal of topic modeling is to create a probabilistic generative model from the corpus of text documents. For that purpose, two state-of-art probabilistic clustering algorithms are considered: Latent Dirichlet Allocation (LDA) [Blei et al. 2003] and Non-negative Matrix Factorization (NMF) [Pauca et al. 2004].

LDA is a probabilistic unsupervised algorithm based on the Dirichlet prior on mixture weights of topics for each text. Its formulation assumes that documents can be represented as a random mixture of latent topics, in which each topic is associated to a probability distribution in relation to the text words. On the other hand, NMF refers to a group of algorithms to approximate a matrix by two matrices, so that the weights defined in one matrix and the components defined in the other matrix are always non-negative. These algorithms are often used in circumstances in which there is no meaning for negative values [Gillis 2014].

3.5. Sentiment analysis

The following steps receive as input the reduced space obtained from a dimensionality reduction algorithm (PCA or t-SNE) associated to the TF-IDF vector space. The three tasks considered in the proposed methodology are described below.

3.5.1. Visual analytics

Visual analytics processes are powerful strategies for implicit knowledge discovery on large and high dimensional datasets, since it incorporates data visualization into analy-

sis tasks [Paulovich et al. 2007]. Basically, visualization techniques generates intuitive graphical representations (also called layouts) of multidimensional data by taking advantage of the human perception system.

A particular case of visualization techniques are those based on a point placement strategy, which operate by placing points in a visual space (1D, 2D or 3D) representing individual data instances. The output is a layout, defined by a set of points in the visual space, in which the proximity of such points reflects some relationship (e.g. similarity) in the original high dimensional space. For that purpose, dimensionality reduction can be employed to perform the mapping of high dimensional data points to a lower dimension. Therefore, PCA and t-SNE are considered for reducing the dimensionality of the vector-space model, defined by the TF-IDF feature vectors.

3.5.2. Clustering

Clustering are fundamental machine learning techniques for grouping data instances by taking into account only its similarity relationships. Among the clustering techniques in literature, K-Means is a state-of-art algorithm that partitions data instances in K clusters, which are represented by centroids.

In the proposed methodology, K-Means is applied to determine clusters in the tweet set and to label them according to the predefined categories. Such procedure is conducted separately, since it does not support multi-label prediction. In K-Means, the number of centroids K is adjusted by means of experiments, while the centroids' values are started at random and L2 norm is used to compare the similarity among data instances.

3.5.3. Classification

Sentiment classification is a relevant topic that aims at identifying the category (polarity, topic or sentiment) of a text. In this sense, as this task is relevant for predicting stock market scenarios and the behavior of investors, the proposed methodology incorporates two sentiment classification approaches. As previously mentioned, each instance may be labeled by more than one sentiment, allowing us to adopt a multi-label classification and to consider the single label classification.

The classification models employed in our methodology are Support Vector Machine (SVM) and Random Forest. We chose such models due to the natural support to multi-label classification and its previous successful employments in literature for text mining and sentiment classification. The next section details parameter tuning of the machine learning algorithms.

4. Experimental results

The proposed methodology was implemented in Python, supported by the sklearn library [3]. The default values of the classifier parameters were kept for the experiments, since no significant performance improvement was observed when modifying these parameters. The analyzed metrics for classification performance were Precision, Recall and F1-Score.

The goal of the first experiment is to determine the optimal K for the K-Means

[3]scikit-learn Machine Learning in Python `https://scikit-learn.org/stable/`

algorithm, which is achieved by the elbow curve test. In this test, K-Means is performed by considering several values for K and verifying the abrupt change in the distortion scores, i.e., the values of the K-Means cost functions. According to the elbow curve, the number of clusters considered by K-Means is set to three, since the distortion reduces significantly when the values for K are greater or equal to three.

4.1. Topic Extraction and Analysis

According to the previous experiment, the appropriate value $K = 3$ is also considered for topic extraction from the tweets. The key strategy is to explore the relations between tweets, groups and extracted topics, as well as to analyze the similarity relationships among the detected clusters.

The number of topics extracted by the LDA and NMF algorithms took into account three clusters. The number of words in the topics extracted was set to 10, since no relevant information was revealed for topic titles longer than 10 words. The number of TF-IDF features was set to 5292, which is the number of unique words in the instances, after the preprocessing step.

The topics extracted by LDA are the following:

- petr4 2014 financeiras 2013 demonstracoes petrobras não fundo cruz3 dilma
- petr4 indicação vale5 confira ações kgt1yitbf7 última resultou suportes volume
- petr4 vale5 gráfico ações diário romper analise trade rastreamento live

The topics extracted by NMF are described as:

- indicação resultou última kgt1yitbf7 confira mrve3 itub4 krot3 oibr4 usim5
- gráfico diário rastreamento analise romper ações 16h 14h 12h brfs3
- petr4 suportes resistências vale5 intraday petrobras não petr3 ibov trade

It can be seen that the topics extracted by both methods did not immediately reveal a correlation with the sentiments attributed to the tweets.

The topics are related to the markets scenario at the time when these tweets were posted at the social network. The terms “petr3” and “petr4” are the stock tickers for the Brazilian company “Petrobras”, which was amongst the largest Brazilian companies in terms of market capitalization in the years of 2013 and 2014. At the time, the company “Petrobras” was being commented a lot in the media due to the raising suspicious activities for political corruption purposes. Furthermore, some topics are related to other companies in the Brazilian Stock Market, since they contain terms representing stock tickers for various companies and describing day-trading operations.

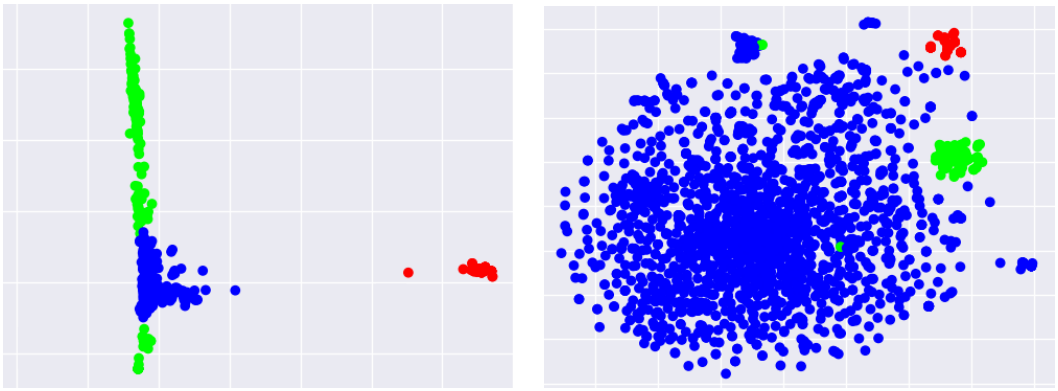
4.2. Sentiment Clustering and Visual Analytics

Figures 4.2(a) and 4.2(b) depicts the layouts obtained from the visualizations of the dataset using PCA and t-SNE, respectively. Each point is associated to a specific tweet and its color is related to a cluster after performing K-Means ($K = 3$): cluster 1 (blue), cluster 2 (red) and cluster 3 (green).

The layouts produced by both visualizations present different clusters geometries and point distribution in the visual space. t-SNE shows rounded clusters of different sizes and densities, in which the similarity relations of cluster 1 were not preserved due to its heterogeneous pattern in the layout. PCA’s layout presents clusters with distinct geometries, but their densities indicate the preservation of similarity relations of instances

within clusters 1 and 2. For that reason, we adopt the low dimensional space produced by PCA from the TF-IDF vectors to the sentiment classification task.

Furthermore, a visual inspection on these clusters reveals that the associated words are similar to the topics extracted by LDA and NMF, but presenting similarity relations to the topics extracted by NMF. The topics extracted by LDA were considerably influenced by "petr4" and "vale5", which are the stocks that contribute the most to the *Ibovespa* index.



(a) (b)

Figure 2. Layouts: (a) K-Means applied to the points on the reduced space obtained by PCA; (b) K-Means applied to the t-SNE reduced data.

The top 10 words per cluster when applying the K-Means algorithm are shown on Table 1. It can be seen that the clusters are related to the topics extracted by NMF. Clusters 1 and 2 present relations to "Petrobras" and day-trading operations. Cluster 3 denotes specific activities from other companies in the Brazilian Stock Market, such as indications, results and consequences of the time when tweets were collected.

Table 1. Top words by cluster.

Cluster 1	Cluster 2	Cluster 3
petr4	resistências	indicação
vale5	suportes	resultou
ações	intraday	última
petrobras	vale5	kgt1yitbf7
gráfico	petr4	confira
diário	bbdc4	mrve3
não	itub4	krot3
trade	elpl4	itub4
vale	comentário	oibr4
2014	sbsp3	usim5

4.3. Sentiment Classification

The sentiment classification step considered multi-label and single-label approaches. As previously mentioned, the classifiers SVM and Random Forest were adopted to classify

tweets according to specific sentiments. The dataset of all tweets was split into 70% training set and 30% test set. The labels were fit to multi-label classifiers. Tweets containing a specific sentiment are labeled as one-valued, while absent sentiments to the tweet are labeled as zero-valued.

Furthermore, in the single-label approach, the classifiers were fit to the data for each category individually and the sentiment prediction was performed for one category at a time.

Table 2. Precision, Recall and F1-Score for Multi-Label SVM

Sentiment	Precision	Recall	F1-Score
TRU	0.42	0.25	0.32
DIS	0.57	0.35	0.44
JOY	0.46	0.24	0.31
SAD	0.34	0.20	0.25
ANT	0.43	0.26	0.32
SUR	0.28	0.15	0.20
ANG	0.59	0.41	0.48
FEA	0.20	0.03	0.05
NEUTRAL	**0.61**	**0.48**	**0.54**

For comparison purposes, the classification performance was compared using the dataset under its high dimensional form (TF-IDF space model) and its low dimensional form, obtained by PCA algorithm. Using the dimensionality reduction, the classifiers SVM and RNF were not able to correctly classify the instances for the considered classes, except for the neutral class. However, the F1-Score for this case was 0.18 for the presence of the neutral class. As the classification performance using the low dimensional data are similar to the high dimensional dataset, the following tables contain only results when using the TF-IDF vector.

The results for these algorithms are described on Tables 4.3 and 3. The numbers in bold represent the best results for the metric between all the categories.

Table 3. Precision, Recall and F1-Score for Multi-Label Random Forest

Sentiment	Precision	Recall	F1-Score
TRU	0.38	0.14	0.20
DIS	0.62	0.26	0.37
JOY	0.45	0.16	0.24
SAD	0.28	0.08	0.12
ANT	0.27	0.08	0.12
SUR	0.18	0.05	0.08
ANG	**0.71**	0.26	0.37
FEA	0.00	0.00	0.00
NEUTRAL	0.59	**0.40**	**0.48**

The individual category classification are described on Table 4. The numbers highlighted in bold represent the best result for each metric in each class.

Table 4. Metrics for classification applying multiple algorithms

Sentiment		Precision		Recall		F1-Score	
		SVM	RF	SVM	RF	SVM	RF
TRU	0	0.72	**0.77**	0.88	**0.89**	0.79	**0.83**
	1	**0.48**	0.39	0.24	0.20	0.32	0.27
DIS	0	0.86	**0.87**	0.96	**0.98**	0.91	**0.92**
	1	0.57	**0.61**	0.27	0.20	**0.37**	0.30
JOY	0	0.85	**0.87**	**0.95**	**0.95**	**0.90**	**0.90**
	1	**0.48**	0.21	0.23	0.09	**0.31**	0.12
SAD	0	**0.90**	**0.90**	**0.97**	**0.97**	**0.93**	**0.93**
	1	**0.31**	0.26	0.12	0.08	**0.18**	0.12
ANT	0	0.80	**0.83**	**0.93**	0.92	0.86	**0.87**
	1	**0.44**	0.35	0.19	0.18	0.26	0.24
SUR	0	**0.87**	**0.87**	0.94	**0.97**	0.90	**0.92**
	1	**0.27**	0.11	0.13	0.03	0.18	0.04
ANG	**0**	**0.94**	**0.94**	0.98	**1.00**	0.96	**0.97**
	1	0.57	**0.75**	0.33	0.14	**0.42**	0.24
FEA	0	**0.96**	0.95	**0.99**	**0.99**	**0.98**	0.97
	1	**0.29**	0.17	0.08	0.03	**0.12**	0.06
NEUTRAL	0	**0.74**	0.66	0.71	**0.80**	**0.73**	0.72
	1	0.55	**0.64**	0.58	0.46	0.57	0.53

4.4. Discussion

The results for all the multi-label approaches did not have acceptable values for precision, recall and F1-Score. This is specially due to the unbalance in the dataset. In this sense, experiments disregarding the neutral sentiment were not performed, since the amount of tweets labeled as neutral is significantly higher than the amount of tweets presenting sentiments. The number of instances that do not have a particular sentiment is much higher than those which are labeled to a particular sentiment. The number of true positives for most cases is considerably low when compared to the number of false negatives, which indicates that there may not be defining characteristics for each sentiment, thus making classification harder. Since there are much more instances which are not labeled to a sentiment, the algorithms predict "0", resulting in a low number of true positives and a high number of false negatives.

Even though the classification performance has been improved when the categories were analyzed individually, in some cases the algorithms could not map an instance to a certain class. However, the recall for the class "1" which indicates the presence of the sentiment in the instance is still low in every case, which means that the algorithm could not obtain the relevant instances for classification, and the number of true positives for the presence of the sentiment was low.

The clusters found by the K-Means algorithm were not directly related to the sentiments involved in the classification, but they were similar to the determined topics. The algorithms may have performed better if the considered dataset would present more instances for each sentiment.

The topics and the clusters found were related, indicating that the algorithms were able to identify common characteristics between tweets. There are tweets about the com-

pany “Petrobras”, tweets about other companies and tweets about day-trading.

This method is useful to identify important general information about datasets, because it is able to successfully find similar characteristics between tweets. This method is applicable to obtain a general understanding of the dataset. Therefore it could be used to determine the general sentiment for a particular stock or set of stocks.

5. Conclusion

This paper described a methodology to analyze sentiments and to recognize patterns from tweets regarding the Brazilian stock market. The first step filters the relevant words on tweets and computes TF-IDF feature vectors, so that they can be properly used as input to the machine learning approaches. In supervised tasks, the proposed methodology can classify tweets according to many types of sentiments. As unsupervised tasks, the K-Means algorithm and topic modeling techniques (Latent Dirichlet Allocation and Matrix Factorization) can support the knowledge discovery on tweets, thus revealing the similarity relations and relevant patterns amongst clusters, sentiments and extracted topics. A visual analysis process using PCA and t-SNE was also considered for producing intuitive graphical representation of the tweet dataset.

Experimental results indicated that the proposed sentiment classification is capable to predict sentiments of tweets related to the stock market in the Portuguese language. As the dataset used is unbalanced, the algorithms could not find defining characteristics for each sentiment. However, the classification performed better on the absence of sentiments. The clusters determined by K-Means did not reveal patterns in relation to the sentiments, at first. An alternative is to explore if the dataset is divided into clusters that truly represent the sentiments.

Future work can be guided on the improvement of the sentiment classification results, since the prediction of stock prices is also a relevant topic. Larger labeled datasets of tweets in Portuguese may be required. Moreover, the presence of numbers in tweets affects clustering and topic modeling approaches, because they usually represent relevant words in the tweets. Thus, it could be interesting to explore whether the classification can be improved if such information is removed from the tweets.

References

[Altrabsheh et al. 2014] Altrabsheh, N., Cocea, M., and Fallahkhair, S. (2014). Sentiment analysis: towards a tool for analysing real-time students feedback. In *IEEE 26th International Conference on Tools with Artificial Intelligence*, pages 419–423. IEEE.

[Barberá and Rivero 2015] Barberá, P. and Rivero, G. (2015). Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33(6):712–729.

[Blei et al. 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

[Bollen and Mao 2011] Bollen, J. and Mao, H. (2011). Twitter mood as a stock market predictor. *Computer*, 44:91–94.

[Clark et al. 2018] Clark, E. M., James, T., Jones, C. A., Alapati, A., Ukandu, P., Danforth, C. M., and Dodds, P. S. (2018). A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter. *arXiv preprint arXiv:1805.09959*.

[Galdi and Gonçalves 2018] Galdi, F. C. and Gonçalves, A. M. (2018). Pessimism and uncertainty of the news and investor behavior in brazil. *RAE-Revista de Administração de Empresas (Journal of Business Management)*, 58(2):130–148.

[Giachanou and Crestani 2016] Giachanou, A. and Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28.

[Gillis 2014] Gillis, N. (2014). The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12.

[Jolliffe 2011] Jolliffe, I. (2011). *Principal component analysis*. Springer.

[Li et al. 2016] Li, H., Cui, J., Shen, B., and Ma, J. (2016). An intelligent movie recommendation system through group-level sentiment analysis in microblogs. *Neurocomputing*, 210:164–173.

[L.Lima et al. 2016] L.Lima, M., P. Nascimento, T., Labidi, S., S. Timbo, N., V. L. Batista, M., N. Neto, G., A. M. Costa, E., and R. S. Sousa, S. (2016). Using sentiment analysis for stock exchange prediction. *International Journal of Artificial Intelligence & Applications*, 7:59–67.

[Maaten and Hinton 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[Pauca et al. 2004] Pauca, V. P., Shahnaz, F., Berry, M. W., and Plemmons, R. J. (2004). Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 452–456. SIAM.

[Paulovich et al. 2007] Paulovich, F. V., Oliveira, M. C. F., and Minghim, R. (2007). The projection explorer: A flexible tool for projection-based multidimensional visualization. In *XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, pages 27–36.

[Plutchik 1980] Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. Harper and Row.

[Rao and Srivastava 2012] Rao, T. and Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 119–123, Washington, DC, USA. IEEE Computer Society.

[Silva 2018] Silva, F. V. d. (2018). Brazilian stock market tweets with emotions.

[Smailović et al. 2013] Smailović, J., Grčar, M., Lavrač, N., and Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In Holzinger, A. and Pasi, G., editors, *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 77–88, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Zimek et al. 2012] Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387.