Development and analysis of a labeled dataset of Brazilian agricultural market news titles for sentiment analysis

Tiago Santa Maria Rodrigues Marto[1], Roberto Fray da Silva[1], Angel Felipe Magnossão de Paula[1], Anna Helena Reali Costa[1], Carlos Eduardo Cugnasca[1]

[1] Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da Universidade de São Paulo, São Paulo, São Paulo, Brasil, tiago.marto@usp.br, roberto.fray.silva@gmail.com, angel.magnossao@gmail.com, anna.reali@usp.br, carlos.cugnasca@usp.br

RESUMO

A análise de sentimentos pode ser definido como o uso de técnicas de processamento de linguagem natural para identificar, extrair e quantificar informação subjetiva e estados emocionais a partir de dados textuais de diferentes origens. No entanto, não existem *datasets* no domínio agrícola de alta qualidade que possuam valores de sentimentos. O principal objetivo deste artigo é criar e analisar, dos pontos de vista estatístico e linguístico, um *dataset* de títulos de notícias dos mercados agrícolas em português. Foram utilizadas três fontes de notícias de mercado confiáveis e relevantes. O *dataset* contém 335 notícias, de 01-03-2019 a 31-05-2019, rotuladas em duas dimensões: valor de sentimento e urgência. A primeira é relacionada ao impacto esperado das notícias no mercado agrícola e contém cinco categorias: muito negativo, negativo, neutro, positivo e muito positivo. A segunda é relacionada à probabilidade de que a notícia em questão impacte no mercado agrícola no dia de sua publicação e contém duas categorias: urgente e não urgente. A maioria das notícias é positiva e não urgente, refletindo a situação do mercado no período em questão. O *dataset* pode ser utilizado para estimar o sentimento do mercado e melhorar tanto previsões de mercado quanto a tomada de decisões.

PALAVRAS-CHAVE: Agricultura, Análise de sentimentos, Inteligência artificial

ABSTRACT

Sentiment analysis can be defined as the use of natural language processing techniques to identify, extract and quantify subjective information and emotional states from textual data from different sources. Nevertheless, there are no high-quality datasets with sentiment values for the agricultural domain. The main objective of this paper is to build and analyze, from the statistical and linguistic points of view, a dataset of agricultural market news titles in Brazilian portuguese. Three trustworthy and relevant market news sources were used. The dataset contains 335 news from 01-03-2019 to 31-05-2019, which were labeled in two dimensions: sentiment value and urgency. The first is related to the expected impact of the news on the agricultural market, and contains five categories: very negative, negative, neutral, positive, and very positive. The second is related to the probability of impacting the agricultural market at the same day that the news was released, and contains two categories: urgent and non-urgent. The majority of the news is positive and non-urgent, reflecting the market situation at the period. This dataset can be used to estimate market sentiment and help on improving market prediction and decision making.

KEYWORDS: Agriculture, Sentiment analysis, Artificial intelligence

INTRODUCTION

The agroindustrial domain comprises all companies involved in activities related to food production, from production at the farm to processing, transportation, warehousing and retailing. The companies, also referred to as agents, and their relations, are the field of study of supply chain management. This considers all companies that are directly or indirectly involved in the fulfillment of a customer's group demands (CHOPRA, MEINDL, 2013).

One important aspect of this domain, that directly influences decision making, is the market sentiment. Sentiment analysis, also referred to as opinion analysis, is the use of natural language processing techniques to identify, extract and quantify subjective information and emotional states from textual data from different sources, such as news, message boards, social media, among others (LIU, 2012; AGGARWAL, ZHAI, 2012). Several domains explores these techniques, such as: finance, using sentiment analysis to evaluate the market sentiment (REN, WU, LIU, 2018; NASSIRTOUSSI et. al, 2014); product reviews, to evaluate how customers rank different products (MUKHERJEE, BHATTACARYYA, 2012); movie reviews, to help to predict the box office (PANG, LEE, 2004); among others.

Sentiment analysis applied on agricultural market news could help to better understand what is the current sentiment on the agroindustrial domain, providing useful information for decision making. Nevertheless, there are almost no papers that explore sentiment analysis on the agroindustrial domain. One such example is Froehlich et. al (2017), which evaluates the worldwide public perception in the aquaculture sector. This lack of papers is even more present on the agroindustrial domain in Brazil.

One of the reasons that explain this lack of papers is the fact that are, to the best of our knowledge, no open datasets for agricultural market news in Brazil. Without a labeled, high-quality dataset, it is not possible to use sentiment analysis to estimate market sentiment. Our objective in this paper is to develop and analyze a labeled dataset with agricultural market news, which can be used for sentiment analysis on this domain.

The news were gathered from three relevant and trustworthy news sources (Valor Econômico, UOL Economia and InfoMoney), and were labeled independently by three researchers on two dimensions: sentiment value and urgency. The first dimension is related to five sentiment classes (very negative, negative, neutral, positive and very positive), as several sentiment analysis papers show that only three sentiment classes (negative, neutral and positive), as used in most of the literature, may not be enough to capture all the sentiments in a text (BOLLEN, MAO, ZENG, 2011; NASSIRTOUSSI et. al, 2014).

The second dimension is related to the subjective probability that the specific news will impact on the agricultural market on the same day of its release. Even though our dataset can be used for both regression and classification problems, because it provides the individual sentiment scores as well as the aggregated ones, we believe it will provide better results on classification tasks. Finally, we conduct an exploratory analysis of the dataset, considering both statistical and natural language processing aspects, and provide suggestions for its use.

METHODOLOGY

The methodology used in this paper contained five main steps:

1. **Identification of the main market news sources that provide publicly available news on the agricultural domain.** Three sources were identified: Valor Econômico (https://www.valor.com.br/), InfoMoney (https://www.infomoney.com.br/) and UOL Economia (https://economia.uol.com.br/). These are considered trustworthy and among the most relevant news sources for financial and market information;

2. **Data collection.** First, we developed multiple web crawlers with the scrapy library (https://scrapy.org/), considering the different websites and their sections. Then, we used them to gather agricultural market news titles from the sources identified in Step 1, from 01-03-2019 to 31-05-2019. It is important to observe that only news that were publicly available were gathered;
3. **Data preprocessing.** In this step, we evaluated the dataset and removed incorrect, irrelevant and redundant news. The main criteria used for identifying irrelevant news was: spams and news that had not even a remote connection with agriculture and the agricultural markets. The main library used in this Step was Pandas (https://pandas.pydata.org/);
4. **Data labeling,** using Pandas and Google Sheets (https://docs.google.com/spreadsheets/u/0/). In this Step, we labeled the data on two main dimensions: (i) urgency, related to the probability that the specific news would impact on the market at the day of its release; and (ii) sentiment value, considering the most probable impact on the market of the specific news. The first one consisted of two classes (high and low probability). The second one consisted of five categories (very negative, negative, neutral, positive, and very positive). Every data point was independently labeled on the two dimensions by three researchers. Then, the resulting labels for each data point in each dimension were calculated. For urgency, the final label was the majority of the labels for that data point. For sentiment value, a rounded up simple average for the labels of the data point was used**;**
5. **Exploratory data analysis.** This step consisted of an analysis of the distribution of the labels on the whole dataset and aggregated by day, using a simple average of the data points of that day. We also analyzed the content of the news titles, in terms of: average number of words, most important words, and words that are specific for the agricultural domain, which may impact on the sentiment analysis task. The main libraries used on this Step were: Pandas, Matplotlib (https://matplotlib.org/), Spacy (https://spacy.io/) and Seaborn (https://seaborn.pydata.org/). The results of this step will be described in the Results Section.

RESULTS

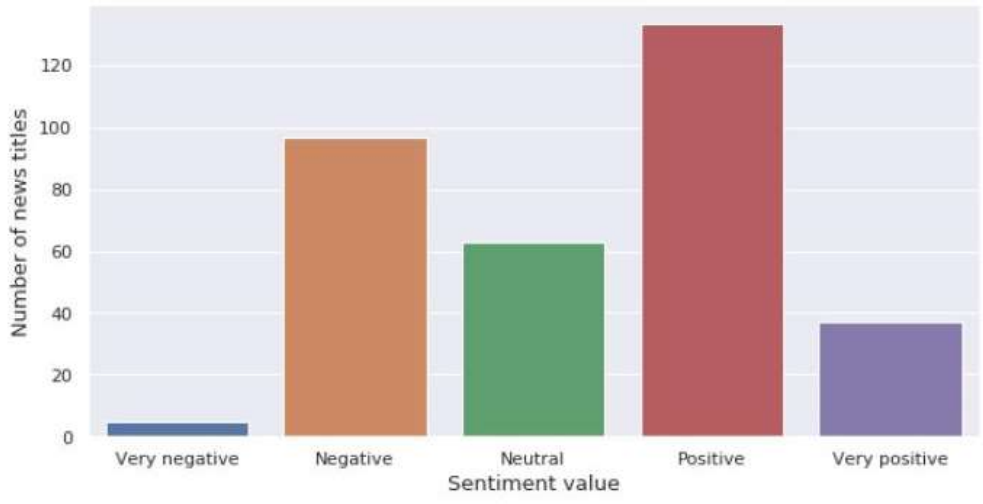
Exploratory data analysis

The raw data collected using the crawlers resulted in a dataset of 2.134 news titles. These were composed of news titles from different markets, such as the stock, commodities markets (such as agricultural commodities, iron, and oil), and electronics markets. Around 10% of these were spam. After carefully analyzing and cleaning the data, it resulted in a dataset of 335 agricultural markets news titles that are relevant for the domain.

After the data cleaning process, we labeled the data in the dimensions of sentiment value and urgency. Figure 1 and Table 2 illustrate the aggregated results of the sentiment value dimension. It is important to observe that the dataset is considerably positive (around 50% of the total dataset), while only around 1.5% is very negative. This reflects the current market situation, in which the exchange rates are favorable for exporters, impacting positively the agricultural domain. Table 2 also shows an example of a sentence for each category.

As for the urgency dimension, only around 15% of the news titles were considered urgent, meaning that they would have a high probability of impacting the market sentiment on the day of its release. Table 3 shows an example of a sentence for each category in this dimension.

Figure 1. Distribution of the sentiment classes on the dataset



As news sentiment values may vary widely within the same day (for example, on some days there were very positive and very negative news), it is essential to aggregate multiple sentiment values in a given period to correctly interpret the market sentiment. The daily sentiment was calculated considering a simple average of the sentiment values for each

day in the dataset. The same procedure was used for calculating the weekly sentiment. Figure 2 illustrates these two variables. It is possible to observe that, even though the daily sentiment varies considerably from one day to the next, the weekly sentiment shows a stable rising trend with some positive peaks at the end of March and beginning of May.

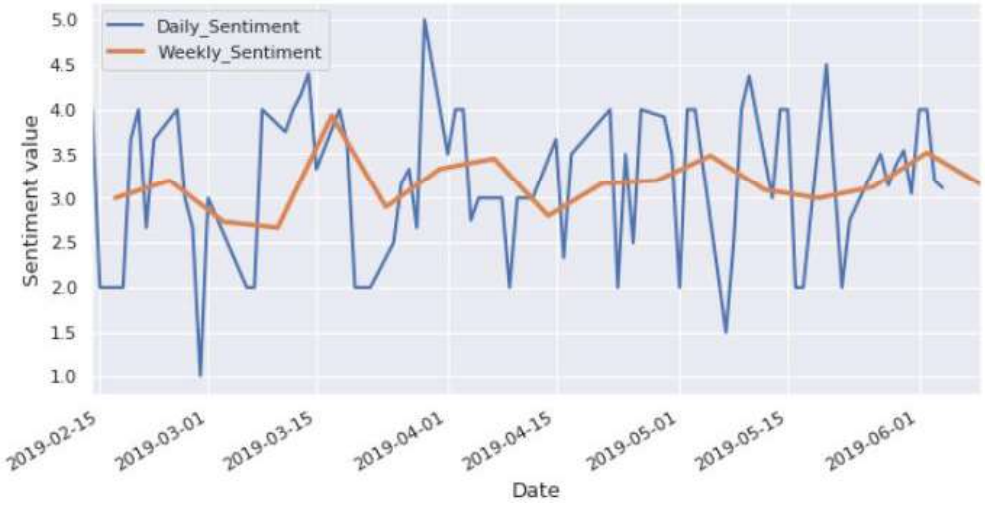
Table 2. Number of data points, percentage and example of sentences for the sentiment dimension

Sentiment class	Number of data points	Percentage of the total dataset	Example of a sentence
Very negative	5	1.49%	BRF loss almost triples in the fourth quarter and totals R$ 2.1 billion
Negative	97	28.97%	Soybeans still under pressure in Chicago and closing prices show a downtrend
Neutral	63	18.83%	CRAs emissions totaled R$ 4 billion in the first quarter
Good	133	39.72%	Conab and IBGE raise projections for grain harvests
Very good	37	10.99%	Brazilian coffee exports increase by 36% in February
Total	335	100%	

Table 3. Number of data points, percentage and example of sentences for the urgency dimension

Sentiment class	Number of data points	Percentage of the total dataset	Example of a sentence
Urgent	52	15.52%	Brazilian coffee exports grow 123% in May, Secex reveals
Non-urgent	283	84.48%	Paraná could lose second place in soybean production to the Rio Grande do Sul
Total	335	100%	

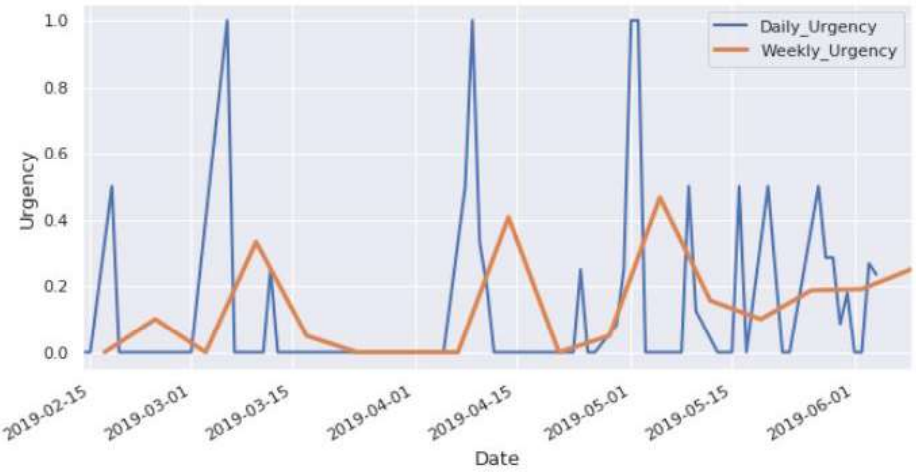
Figure 2. Daily and weekly sentiments of the news on the dataset.



These results reflect the reality: the market was indeed increasing its expectations from the beginning of 2019 up to the end of the dataset. The differing volatility between daily and weekly sentiments is also observed in the financial market, and we believe that our results reflect the reality of the agricultural markets. With a larger dataset, the weekly and monthly sentiment could be used to point the overall trend in sentiment value, and the daily sentiment could be used for speculation purposes.

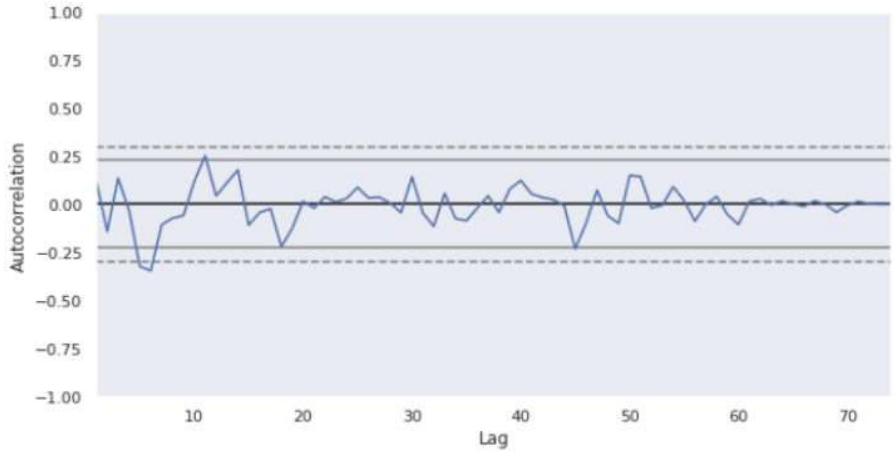
Figure 3 illustrates the daily and weekly urgencies of the news on the dataset. Some partial conclusions that can be drawn from this chart (as more data is needed in this dimension to clearly evaluate the frequency of urgent news related to non-urgent ones) are that: (i) it does present some correlation with the peaks and valleys of sentiment value; and (ii) that it may be influenced by the sentiment volatility in the market (or itself influence this volatility). These results are interesting because this dimension could also be an input for decision making, as well as the sentiment values, especially if the agent wants to better identify the periods of higher volatility in the market sentiment.

Figure 3. Daily and weekly urgency of the news on the dataset.



An analysis of autocorrelation of these two dimensions (sentiment value and dimension) on both daily and weekly frequencies showed that there is autocorrelation on daily sentiment. This is illustrated in Figure 4. Nevertheless, more data is needed to make a thorough evaluation of autocorrelation. It is not possible to discard, at the moment, if weekly sentiment or the daily and weekly urgencies are not autocorrelated variables.

Figure 4. Autocorrelation plot for Daily Sentiment.



A brief analysis of the linguistic properties of the dataset showed that it contains 773 unique words, from which 95 are considered stopwords for the Portuguese language. This results in 678 relevant words that can be used for sentiment analysis. Nevertheless, a considerable portion of these is not relevant for the domain, such as: dirigente (leader), days of the week, month names, and location identifications (such as south, center-west, etc), among others.

An analysis of the morphology of the dataset shows that the most important classes, in terms of frequency of words, are: verbs, propositions, numerals, and nouns. An analysis of the entities on the sentences, using Spacy's named entities recognition module (https://spacy.io/usage/linguistic-features#named-entities), showed some interesting results: (i) only very large multinational companies have their entities on their dictionary (the main examples are Marfrig and BRF, very important food companies); and (ii) countries are correctly identified (such as the USA and China).

DISCUSSIONS

One of the biggest barriers for conducting sentiment analysis on the agricultural market, as pointed before, is the lack of datasets with sentiment values. In this paper, we developed a labeled dataset, considering two dimensions: sentiment values and urgency. We also conducted statistical and linguistic analysis.

There were two main difficulties in conducting this research: (i) creating the dataset itself, which involved choosing trustworthy sources and gathering data from them; and (ii) a lack of a well-established methodology for labeling the dataset. We believe that the methodology used can be used for generating datasets for other domains, as the results were

satisfactory. One important aspect to consider is the necessity to label the dataset using additional dimensions, what depends on domain-specific characteristics.

Although more data is needed to create robust sentiment prediction models, we are currently conducting experiments with several machine learning models, such as artificial neural networks, long short-term memory networks, and support vector machines. Future work is needed on improving the dataset, including more data points, further refining the sentiment value and urgency dimensions, and also implementing sentiment analysis and prediction models. We believe that the results of these researches could improve financial prediction models applied to the agricultural markets, improving decision making.

It is very important to observe that, as a considerable portion of the unique identified words in the dataset are domain-specific (related to agricultural processes, products, or simply the relation between market and environmental variables and the quantity and quality of products on a given season), we can infer that a general lexicon with sentiment values would not be able to correctly identify and predict sentiments in this domain.

CONCLUSIONS

Sentiment analysis techniques could help improving decision making in the agricultural sector. Nevertheless, there are no few papers that conduct this type of research on the Brazilian agricultural market, and there are no datasets that contain sentiment values available for research. The main objective of this paper was to develop and analyze a dataset of agricultural market news titles that could be used for sentiment analysis on this domain, as well as analyze it from the statistical and linguistic points of view.

The dataset developed contains 335 news titles, labeled in two dimensions: (i) sentiment value, which contains five categories (very negative, negative, neutral, positive, and very positive); and (ii) urgency, which contains two categories (urgent and non-urgent). On the first dimension, it contains mostly positive news. On the second dimension, it contains mainly non-urgent news.

Through the analyses that were conducted, we conclude that the dataset is suitable for sentiment analysis tasks, which will be developed on further researches. The methodology used can be applied to develop labeled datasets for other domains, improving the sentiment analysis and prediction in those domains as well.

ACKNOWLEDGEMENTS

This work was supported by Itaú Unibanco S.A. through the Itaú Scholarship Program, at the Centro de Ciência de Dados (C2D), Universidade de São Paulo, Brazil.

REFERENCES

AGGARWAL, C.C.; ZHAI, C. Mining text data. Springer Science & Business Media, 2012, 527 pp.

BOLLEN, J.; MAO, H; ZENG, X. Twitter mood predicts the stock market. Journal of Computational Science, v. 2, n. 1, p.1-8, 2011.

CHOPRA, S.; MEINDL, P. Supply chain management: Strategy, planning, and operation, 5th ed., New Jersey, USA: Pearson Education, 2013, 528pp.

FROEHLICH, H.E.; GENTRY, R.R.; RUST, M.B.; GRIMM, D.; HALPERN, B.S. Public perceptions of aquaculture: evaluating spatiotemporal patterns of sentiment around the world. PloS one, v. 12, n. 1, p.e0169281, 2017.

HUTTO, C.J; GILBERT, E.E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, 2014.

LIU, B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, v. 5, n. 1, p.1-167, 2012.

MUKHERJEE, S.; BHATTACHARYYA, P. Feature specific sentiment analysis for product reviews. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Berlin, Heidelberg, p. 475-487, 2012.

NASSIRTOUSSI, A.K.; AGHABOZORGI, S.; WAH, T.Y.; NGO, D.C.L. Text mining for market prediction: A systematic review. Expert Systems with Applications, v. 41, n. 16, p.7653-7670, 2014.

PANG, B.; LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

REN, R.; WU, D.D.; LIU, T. Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Systems Journal, v. 99, p.1-11, 2018.

SVM com DenseNet para classificação de mudas do guaraná a partir da nervura central do folíolo

Brenda Talyne Costa Martins[1], *Marcos Filipe Alves Salame*[2]

[1]Bolsista de Iniciação Científica FAPEAM, Embrapa Amazônia Ocidental
Manaus, Amazonas, Brasil, brendatalyne@gmail.com

[2]Analista de Tecnologia da Informação, Embrapa Amazônia Ocidental
Manaus, Amazonas, Brasil, marcos.salame@embrapa.br

RESUMO

O guaranazeiro *(Paullinia cupana var. sorbilis)* é um importante e tradicional cultivo do Estado do Amazonas. Com o aumento da demanda por guaraná, a Embrapa Amazônia Ocidental desenvolveu e registrou 20 cultivares e lançou 18 até o momento, com o objetivo de aumentar a produtividade, reduzir o tempo de formação, fornecer alto teor de cafeína e mais resistência às principais pragas. Entretanto, o procedimento de identificação desses cultivares é difícil e feito somente por especialistas. Diante deste cenário, foi criado um *dataset* de mudas de guaraná, composto por 10 cultivares, através de um protótipo desenvolvido para dispositivos móveis, de forma a evitar a necessidade de qualquer pré-processamento manual e visando a proximidade com o ambiente real do produtor. Foram realizados experimentos com 5 dos cultivares coletados, utilizando aprendizado de máquina com SVM e DenseNet para classificação a partir de imagens da nervura central dos folíolos e obteve-se como resultado uma acurácia de 88%.

PALAVRAS-CHAVE: Aprendizado de Máquina, guaranazeiro, cultivares.

ABSTRACT

The guarana plant (Paullinia cupana var. Sorbilis) is an important and traditional crop of state of Amazonas. Due to increase of demand for guarana, Embrapa Amazônia Ocidental developed and registered 20 cultivars and launched 18 so far, aiming to increase productivity, reduce formation time, provide high caffeine content and more resistance to major pests. However, the identification procedure of these cultivars is difficult and it is made only by specialists. Based on this scenario, we created a dataset of guarana seedlings, composed by 10 cultivars, through a prototype developed for mobile devices, in order to avoid the need for any manual preprocessing and aiming at proximity to the environment of the rural producer. Experiments