Sequential Short-Text Classification from Multiple Textual Representations with Weak Supervision

Ivan J. Reis Filho[1,2][0000−0003−1968−6279], Luiz H. D. Martins[1][0000−0003−2904−1896], Antonio R. S. Parmezan[2][0000−0002−1725−132X], Ricardo M. Marcacini[2][0000−0002−2309−3487], and Solange O. Rezende[2][0000−0002−5233−7639]

[1] Minas Gerais State University, Frutal, Brazil
[2] University of São Paulo, São Carlos, Brazil

Abstract. The amount of news generated on the internet has increased significantly in recent years. As a trend, text data has gained attention from industry, government, academia, and the financial market. This information is potentially valuable to assist domain experts in decision making. Therefore, related applications based on machine learning have been widely available in several areas of knowledge. However, for supervised learning tasks, the availability of annotated texts in quantity and quality is a recurring problem. This work proposes a time-series-driven approach to labeling chronologically arranged documents. Our proposal categorizes short texts for a particular domain according to the level and trend patterns of a given time series. We use the obtained weak labels with the understanding that they are imperfect but still useful for building predictive text models. Documents and agribusiness commodity price series were employed to assess performance in four classification scenarios. The experimental evaluation considered nine textual representations and different learning paradigms. Neural language-based models demonstrated better classification performance than traditional ones. The results indicate that the proposed approach can be an alternative for automatically labeling a large news volume.

Keywords: Data Labeling, Machine learning, Text Mining, Weak Supervision.

1 Introduction

We have witnessed an increased interest in machine learning-based applications for industry, government, academia, and the financial market [18]. Supervised learning tasks such as classification and regression have been widely explored to assist domain experts in decision making. In this way, emerging intelligent technologies have enhanced the offer of computational resources capable of storing, analyzing, and predicting information from a large volume of data [7].

Predictive models are generally learned from a dataset containing many training samples, each corresponding to an object or event. In this context, the performance of machine learning models depends on the availability of labeled data in sufficient quantity and quality [6]. However, annotated data for some domains can be scarce, and the typical process of obtaining labels with experts inspecting individual samples is usually expensive and time-consuming. Thus, to overcome this limitation, machine learning techniques should be able to work under weak supervision [27].

Weak supervision provides a significantly inexpensive alternative to traditional annotation, reducing the need for humans to hand label large datasets to train machine learning models [8, 6, 25]. Researchers have employed this technique to support many applications, including annotating and detecting fake news [12, 21, 25], labeling images from social media posts [9], recognizing named entities [16], and classifying texts using external sources [8, 18].

In recent decades, the amount of news generated and made available on the internet has grown exponentially [14]. Text mining and natural language processing methods allowed the conversion of such documents into helpful information for experts in different domains [11]. However, due to the lack of annotated news, unsupervised and semi-supervised learning tasks have been adopted for these applications [24]. In light of this, this paper proposes a time-series-driven approach to labeling chronologically arranged documents. Time series are ordered sequences of numerical observations recorded over time. In finance, the price series represents daily records of prices practiced on the stock exchange or commodities. Sudden fluctuations in the price series can mean political, climatic and macro-economic events, as well as market supply and demand.

Interestingly, events that alter market behavior are often reported explicitly in text news. Thus, we design in this work a function that uses the price series of two Brazilian agribusiness commodities to label short texts that correspond to agricultural news. Our proposal weakly categorizes the documents according to the time series's level and trend patterns. We use the obtained weak labels with the understanding that they are imperfect but still useful for building predictive text models. An experimental evaluation estimated the efficiency of our approach in the face of nine textual representations and different learning paradigms. Furthermore, we propose a vector representation of texts based on bag-of-words that uses a distance measure between Terms and Documents through pre-trained BERT models, designated here as TD-BERT.

The remainder of this paper is structured as follows: Section 2 describes and contrasts related work. Section 3 presents our contribution. Section 4 reports the empirical evaluation and discusses the results. Finally, Section 5 concludes our study and lists future work.

2 Related Work

There are many strategies for labeling training data automatically. These annotation tactics generate imperfect (less accurate) labels based on domain knowledge

and are commonly pervasive as weak supervision [18]. We can categorize weak supervision approaches into three types [27]: (i) incomplete supervision, where only a small subset of training data is available with annotations; (ii) inexact supervision, in which training data is only provided with coarse annotations; and (iii) inaccurate supervision, where available labels are not always ground-truth.

Many incomplete supervision approaches have been proposed to identify fake news. The studies developed techniques to annotate social media news and increase the amount of training data from various sources [12, 21, 25]. Inexact supervision approaches have been proposed in which some supervisory information is provided but not as accurate as desired [3, 9, 13]. For example, a study [6] developed a framework for weak interactive supervision where a method proposes heuristics and learns from user feedback on each heuristic. The experiments demonstrated that only a few feedback iterations are needed to train models that achieve highly competitive test performance without access to ground-truth training labels.

In this paper, we focus on inaccurate supervision procedures due to the similarity of the proposed approach. Inaccurate supervision concerns the situation in which the supervision information is not always ground-truth, and some annotations may suffer from errors [27]. Weak labeling techniques in text classification tasks are known as distant supervision [17]. Distant supervision generates training annotations by heuristically aligning data points with an external knowledge base [2, 15]. In addition, heuristic rules for labeling data are also common sources of weak supervision. That is, weak supervision sources mainly contain distant supervision [5, 20, 26] and heuristic rules [19, 10].

A study presented a practical approach for treating the identification of fake news on Twitter as a binary machine learning problem [12]. The tweets were labeled by their sources, *i.e.*, tweets issued by accounts known to spread fake news were labeled as fake, and tweets issued by accounts known as trustworthy were labeled as accurate. Two datasets and six textual representation models were considered for experimental evaluation. Two alternatives were explored to represent the tweet textual contents: a Bag-of-Words (BoW) employing TF-IDF vectors and a neural Doc2vec model trained on the corpus. Instead of creating a small but accurate hand-labeled dataset, the authors demonstrated that using a large-scale dataset with inaccurate labels yields competitive results.

A more specific study for short-text classification involving insufficient unlabeled data, data sparsity, and imbalanced classification was reported in [8]. The proposed method can generate probabilistic labels through the conditional independent model. Six pre-training models were adopted: BERT Base and Multilingual Chinese, RoBERTa Base and Large Chinese, ERNIE and ERNIE Chinese. According to experimental results on public and synthetic datasets, unlabeled imbalanced short-text classification problems can be solved effectively by multiple weak supervision. Notably, recall and F_1Score can be improved without reducing precision by adding distant supervision clustering, which can be employed to meet different application needs.

The authors [16] presented an approach to bootstrap named entity recognition models without requiring any labeled data from the target domain. Instead, the approach relies on labeling functions by automatically annotating documents with named entity labels. A Hidden Markov Model (HMM) was trained to unify the noisy labeling functions into a single (probabilistic) annotation, considering each labeling function.

The success of machine learning methods for texts is closely related to the pre-processing strategy of textual data and the characteristics of the application domains [4]. We highlight that studies on textual representations for weak supervision tasks have received much attention in the literature. However, no in-depth studies were found on text classification for supervision heuristically aligning textual and time series data. In this sense, we introduce the proposed method in Section 3.

3 Methods

This work investigates a short-text labeling function of the commodity market using time series data (Fig. 1). In addition, it contemplates a vector text representation model based on BoW that adopts a measure of distance between Terms and Documents from pre-trained BERT models, called TD-BERT. Thus, classification models are applied to assess the predictive performance of the proposed approaches. Fig. 1 illustrates the steps performed in this study.

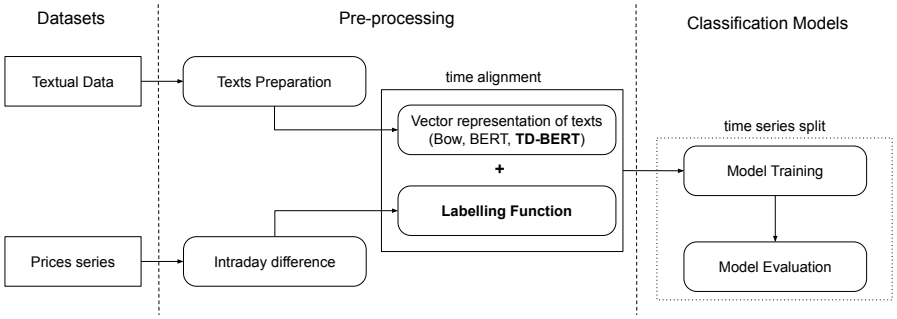

Fig. 1. Conceptual model of the proposed method.

3.1 Labeling Function

A price series S of size m is defined as an ordered sequence of observations, *i.e.*, $S = (s_1, s_2, ..., s_m)$, where s_t represents an observation s at time t. The textual documents D is also an ordered sequence $D = (d_1, d_2, ..., d_k)$, where d_t is a text d at time t, and size n. Therefore, we attribute via time alignment a label (-1, 0 or 1) to texts using the following equation:

$$d_t = \begin{cases} -1 \; if \; s_{t+lag} < (s_t + s_{t-lag})/2 \\ 1 \; if \; s_{t+lag} > (s_t + s_{t-lag})/2 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{1}$$

the text d_t receives a label according to the level and trend patterns of the time series S. The constant lag corresponds to the seasonal period of the time series in number of observations. To exemplify, Fig. 2 portrays the result of Equation 1 applied to a synthetic time series with $lag = 5$. This function aims to capture the time series' stable, increasing, and decreasing behaviors to assign labels to short texts arranged chronologically in time.

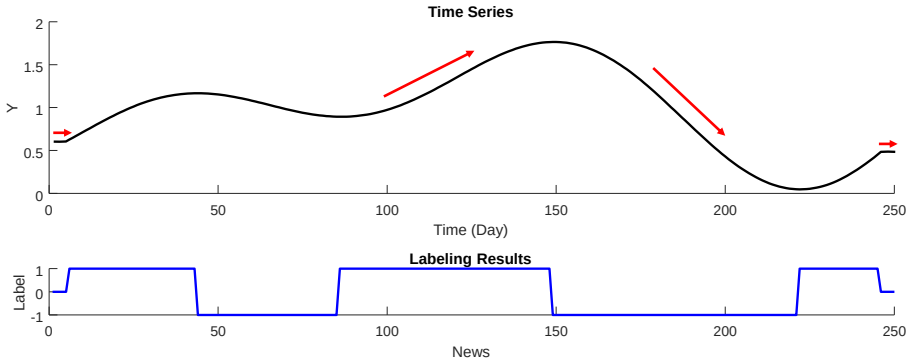

Fig. 2. Illustration of how the labeling function works.

3.2 TD-BERT

We implemented the proposed approach to obtain a new textual representation that considers the semantic features. First, we extract a collection of documents $D = [d_1, d_2, ..., d_k]$ containing k documents and a set $T = [w_1, w_2, ..., w_b]$ with b terms from D. This process is similar to the one used in BoW. However, we take into account here the sentence transformers of the pre-trained BERT models to obtain the cosine distance of each term in each document.

The textual representation D with sentence transformers is defined as $DS = ([B_1], [B_2], ...[B_k])$, where each B is a BERT vector of h positions representing a document d at time t. The representation of Terms with the sentence transformers is defined as $TS = ([W_1], [W_2], ...[W_b])$, where W_j is a BERT vector of h positions that represents a term w_j. The set of documents is represented as a document-term matrix constituted by cosine distance c from each vector k composed of b dimensions, as depicted in Fig. 3.

The matrix values correspond to the cosine distance of each term in each document, *i.e.*, $c(B_k, W_b)$ equals the distance between vectors W_j and B_i. The vector values DS and TS are assigned according to a pre-trained BERT model. Thus, in this work, we evaluate the classification performance applying three

	W_1	W_2	...	W_{b-1}	W_b
B_1	$c(B_1,W_1)$	$c(B_1,W_2)$	...	$c(B_1,W_{b-1})$	$c(B_1,W_b)$
B_2	$c(B_2,W_1)$	$c(B_2,W_2)$	...	$c(B_2,W_{b-1})$	$c(B_2,W_b)$
⋮	⋮	⋮	⋱	⋮	⋮
B_{k-1}	$c(B_{k-1},W_1)$	$c(B_{k-1},W_2)$	...	$c(B_{k-1},W_{b-1})$	$c(B_{k-1},W_b)$
B_k	$c(B_k,W_1)$	$c(B_k,W_2)$	...	$c(B_k,W_{b-1})$	$c(B_k,W_b)$

Fig. 3. Illustration of the representation of document k as a document-term matrix.

pre-trained models: BERT base multilingual (TD-BERT), DistilBERT base multilingual (TD-DistilBERT), and BERT base Portuguese (TD-BERTimbau) [23].

4 Evaluation

We present a weak supervision evaluation of two agricultural commodities datasets: corn and soybean. Furthermore, we compare several predictive models for the classification task, considering distinct textual representations. We selected methods from different machine learning paradigms to better compare the investigated configurations. The K-Nearest Neighbors (KNN) method belongs to the instance-based classification paradigm. Multi-Layer Perceptron (MLP) is an algorithm from the connectionist paradigm. Gaussian Naive Bayes (GNB) and Multinomial Naive Bayes (MNB) are probabilistic methods. Support Vector Machine (SVM), in turn, is a model of the statistical learning theory paradigm.

The following subsections report the steps illustrated in Fig. 1. We present an analysis of the impact of textual representations on the classification of agricultural commodity headlines. These tasks are relevant to emerging research topics related to classifying large volumes of unlabeled text. For reproducibility purposes, we provide a GitHub repository at https://github.com/ivanfilhoreis/ws_text with the source code of the classification methods and the textual representations.

4.1 Datasets

We used texts and time series of corn and soybean. The Portuguese textual data were extracted from an agricultural news website[3]. Founded in 1997, *Notícias Agrícolas* is one of the Brazilian agribusiness's most influential media. Table 1 describes the dataset period, the number of days, and information about the text data.

Time series data were extracted from the Center for Advanced Studies in Applied Economics (CEPEA) at the University of São Paulo (USP).

[3] https://www.noticiasagricolas.com.br/

Table 1. Overview of the time series and textual data used in our experimental evaluation.

Commodity	Corn and Soybean
Period	2015-01-05 to 2021-12-10
Number of Days	1753
TS Attributes	Values (Open, **Close**, High, Low)
Number of headlines/News	7172 (Corn) - 8394 (Soybean)

4.2 Pre-processing

We evaluated the predictive model performances considering weak labels in positive and negative binary scenarios. The Positive Binary (PB) scenario has the labels [0, 1], and Negative Binary (NB) one has the labels [-1, 0]. Table 2 presents examples of labeled headlines in agreement with the function formalized in Section 3.1.

Table 2. News samples labeled using the labeling function.

Com.	**Date**	**Headline**	**lab.**
Corn	2016-01-12	Dólar sobe nesta 4ª com atenção à política interna; milho acompanha	1
	2016-06-21	Preços do milho recuam até 15% no Brasil com colheita da 2ª safra	-1
	2017-03-27	Incerteza sobre a demanda por milho resulta em nova queda de preço	-1
	2018-02-27	USDA reporta a venda de 130 mil toneladas para destinos desconhecidos	1
Soybean	2016-05-10	Chuva do início de outubro ainda não foi suficiente para as lavouras no Sul do MS	1
	2017-01-30	Com queda do dólar e perspectiva de safra elevada, preço da soja cai no Brasil	-1
	2018-11-30	Soja opera estável na Bolsa de Chicago observando início da reunião do G20	-1
	2020-09-21	USDA informa nova venda de 435 mil t de soja para China e demais destinos	1

Among the 7172 corn headlines, 3209 were labeled as negative (-1), 66 as neutral (0), and 3897 as positive (1). Regarding soybean headlines, 3681 were labeled as negative, 82 as neutral, and 4631 as positive. In order to make a binary assessment, negative labels were assigned as neutral in the PB scenario, and in the NB one, positive labels were also changed to neutral.

Our work applies BoW-based representations, pre-trained NLM models, and the proposed TD-BERT model for vector representation of the texts. In the BoW modeling, we used three-term weighting techniques: Binary, TF, and TF-IDF. We considered only unigram versions of each of these weighting terms. In these

models, we applied a text cleaning process to decrease the data dimensionality and increase representation quality. According to [1], this process improves the quality of the classification algorithms. The cleaning steps were: (1) converting words to lowercase and removal of accents; (2) removal of punctuation marks and alphanumeric characters; (3) removal of stopwords; and (4) word-stemming.

We used three pre-trained neural language models to assess weak supervision techniques: Multilingual (M) versions of BERT, M. DistilBERT, and the Portuguese version BERTimbau [22]. In the pre-trained models, we do not use text cleaning techniques to maintain the original text structure, which is essential for context-dependent NLMs. Thus, the sentence transformers of each trained model were employed as input for the predictive models. Also, we used the pre-trained models to build the proposed models: TD-BERT (TD-Be), TD-DilstilBERT (TD-Di), and TD-BERTimbau (TD-Ba).

4.3 Classification Models and Experimental Setup

We used five traditional classification algorithms: MLP, SVM, KNN, GNB, and MNB. The parameters of the ML algorithms we adopted in our experiments were default values of the scikit-learn library.

The time series split evaluation strategy was employed to consider temporal dependence of the textual data, *i.e.*, we train past news to evaluate a future scenario. Thus, seven splits were used for eight evaluations. In this configuration, each split represents one year of the textual dataset. Fig. 4 outlines the time series split assessment strategy adopted in this study.

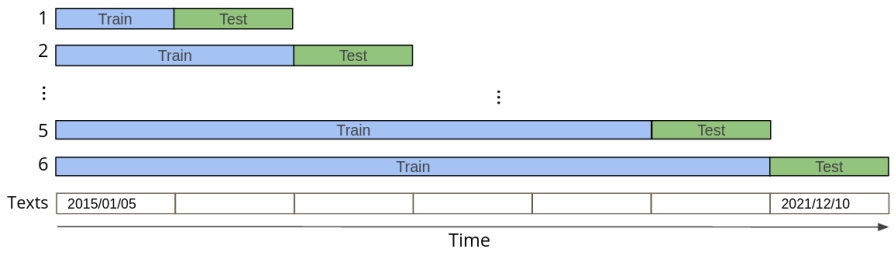

Fig. 4. Time series split used in the experimental setup.

For the evaluation step, we used the F_1 evaluation measure, which corresponds to the harmonic mean of Precision 3 and Recall 4. Equation 2 defines the F_1 index. We employed this metric because the classes are imbalanced in all evaluation splits.

$$F_1 = \frac{2 \times Prec \times Rec}{Prec + Rec}, \tag{2}$$

$$Prec = \frac{TP}{TP + FP}, \tag{3}$$

$$Rec = \frac{TP}{TP + FN}, \tag{4}$$

where TP (True Positive) refers to the number of documents of a class in which the algorithm has correctly classified, and FP (False) indicates the number of documents that do not belong to a class the algorithm wrongly classified as belonging. Finally, FN (False Negative) refers to the number of documents from a class that the algorithm wrongly classified as another class.

4.4 Results and Discussion

We conducted an experimental evaluation to investigate two aspects of weak supervision. In the first aspect, we sought to analyze each textual representation model's impact considering the five different classification algorithms. In the second aspect, we assessed the influence of the neural language model on two weak supervision classification tasks.

Concerning the first aspect, Tables 3 and 5 present the classification results of the MLP, SVM, KNN, GNB, and MNB algorithms on the corn and soybean datasets. This table covers an evaluation scenario PB for Corn and Soybean, which we named CPB and SPB. Each row represents the result of F_1 for a specific algorithm. In bold, we highlighted the highest values for each classification model. The underlined values reflect the best performance of the textual representation models (BoW, BERT and TD-BERT), and the value in parenthesis is the best result considering all the performances.

Table 3. Positive binary evaluation results. Comparison (macro F_1 measure) of BoW models, pre-trained neural language and the proposed TD-BERT hybrid model.

Corn – Positive Binary (CPB)									
Mod.	**Bin.**	**TF**	**TFIDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	0.496	0.495	0.495	0.486	0.488	(**0.499**)	0.378	0.342	0.356
SVM	**0.456**	0.454	0.452	0.431	0.422	0.412	0.416	0.389	0.388
KNN	0.484	0.483	0.490	0.476	0.479	0.492	0.483	0.486	**0.495**
GNB	0.439	0.439	0.444	0.496	0.485	**0.497**	0.494	0.495	0.462
MNB	0.487	0.488	0.451	-	-	-	-	-	-
Soybean – Positive Binary (SPB)									
Mod.	**Bin.**	**TF**	**TFIDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	**0.490**	0.488	0.488	0.476	0.489	0.485	0.344	0.312	0.352
SVM	0.440	0.439	**0.442**	0.398	0.371	0.387	0.381	0.355	0.357
KNN	0.483	0.481	0.484	0.478	0.474	**0.486**	0.477	0.474	0.481
GNB	0.470	0.470	0.469	0.498	0.499	(**0.500**)	0.494	0.481	0.469
MNB	0.472	0.472	0.436	-	-	-	-	-	-

The neural language models were not processed for MNB because it does not accept vectors with negative values. However, we considered it essential to keep the MNB results for the BoW representations in order to compare them with other results. Analyzing the highlighted values of CPB (bold) for each

classification model, we observed that the representations BERTimbau (B.Br), Binary (Bin), and TD-BERTimbau (TD-Br) obtained the best values F_1. We also noticed that the BERTimbau (MLP) model had the highest value among all results (0.499). The SPB results showed Binary, TF-IDF, and BERTImbau as the best values for each F_1 ranking ratio, with BERTimbau (0.500) being the highest value among all the SPB results. Table 4 presents the best CPB and SPB results in terms of precision, recall, and accuracy.

Table 4. Evaluation metrics concerning the best SPB and CPB classification results.

	CPB: BERTimbau (MLP)				**SPB: BERTimbau (GNB)**			
	prec.	recall	f1-score	support	prec.	recall	f1-score	support
0	0.489	0.385	0.422	2917	0.478	0.465	0.458	3358
1	0.534	0.636	0.574	3227	0.544	0.559	0.540	3836
accuracy	0.518				0.511			
macro avg	0.512	0.511	**0.499**	6144	0.511	0.512	**0.500**	7194
weighted avg	0.527	0.518	0.500	6144	0.546	0.511	0.517	7194

The CPB and SPB accuracies were 0.518 and 0.51, respectively. However, looking at the support values of Table 4, we can see that the weak labels are reasonably balanced. Therefore, by analyzing results for this type of evaluation, Macro F_1 becomes more appropriate. Regarding NB, Table 5 displays two assessment scenarios. We called them the Negative Binary classification of Corn (CNB) and Soybean (SNB) approaches. Table 6 lists the best results of CNB and SNB concerning precision, accuracy, and recall.

Table 5. Negative Binary evaluation results. Comparison (macro F_1 measure) of BoW models, pre-trained neural language, and the proposed TD-BERT hybrid model.

Corn – Negative Binary (CNB)									
Mod.	**Bin.**	**TF**	**TFIDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	0.491	0.495	0.496	0.482	**0.497**	0.493	0.358	0.343	0.362
SVM	**0.452**	0.451	0.451	0.428	0.418	0.411	0.406	0.375	0.374
KNN	0.484	0.481	0.490	0.474	0.479	0.492	0.485	0.483	**0.494**
GNB	0.439	0.439	0.443	0.497	0.490	(**0.507**)	0.493	0.494	0.469
MNB	0.491	0.490	0.446	-	-	-	-	-	-
Soybean – Negative Binary (SNB)									
Mod.	**Bin.**	**TF**	**TFIDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	0.48	0.487	**0.488**	0.474	0.481	0.485	0.339	0.288	0.344
SVM	**0.434**	0.432	0.432	0.389	0.363	0.378	0.376	0.358	0.364
KNN	0.471	0.471	0.472	0.471	0.468	**0.480**	0.47	0.47	0.478
GNB	0.451	0.452	0.453	0.500	(**0.501**)	0.496	0.485	0.486	0.461
MNB	0.474	0.473	0.429	0	0	0	0	0	0

Table 6. Evaluation metrics regarding the best CNB and SNB classification results.

	CNB: BERTimbau (GNB)				**SNB: DistilBERT (GNB)**			
	prec.	recall	f1-score	support	prec.	recall	f1-score	support
-1	0.489	0.520	0.492	2881	0.479	0.524	0.483	3339
0	0.554	0.520	0.520	3263	0.553	0.512	0.517	3855
accuracy	0.517				0.506			
macro avg	0.521	0.520	**0.507**	6144	0.516	0.518	**0.501**	7194
weighted avg	0.534	0.517	0.511	6144	0.552	0.506	0.513	7194

Observing the CNB results, we emphasize that the DistilBERT (Diltil.), Binary, TD-BERTimbau, and BERTimbau (B.Br) representations had the highest values (bold) of F_1 for each classification algorithm, respectively. Regarding the SNB results, the TF-IDF, Binary, BERTimbau, and DiltilBERT representations achieved the best results. In this case, the values 0.517 and 0.569, in parentheses, represent the best results of CNB and SNB, respectively.

Aiming to investigate the second aspect of the experimental evaluation, we analyzed the impact of the neural language model on weak supervision. According to the underlined result of CPB and SPB in Table 3, the binary representation had the highest F_1 values, *i.e.*, 0.496 and 0.490, respectively. The Neural language model BERTimbau performed better in the two scenarios with F_1 values of 0.499 and 0.500. Finally, TD-BERTimbau and TD-BERT representations achieved F_1 results with values of 0.495 and 0.494. Thus, representations models based on neural language had better performance than the BoW models. To illustrate the vector distribution of the texts, Fig. 5 presents a graph of the textual representations that performed better in each representation model of Table 3 (underlined values).

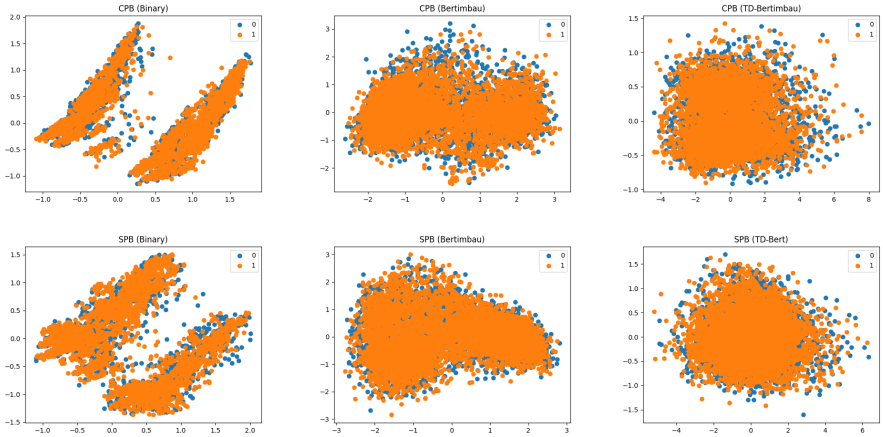

Fig. 5. CPB and SPB. PCA technique for plotting textual representations.

The PCA technique was used to reduce the dimensionality of the textual representation of the agricultural commodities dataset. We observed that headlines classified as positive (1) are more concentrated in the graph distribution, while headlines classified as neutral are a little more sparse. Furthermore, the CPB (TF-IDF) and SPB (TF) representations have smaller ranges on the axes than the BERT-based representations. In this sense, we believe that this broader spectrum can abstract more semantic information from texts.

Comparing the CNB and SNB underlined results in Table 5, the TF-IDF, BERTimbau, DistilBERT, TD-DistilBERT, and TD-BERTimbau obtained the best classification performance for the learning algorithms, respectively. The DistilBERT and BERTimbau neural language models performed better for the GNB method. In both experiments (PB and NB), we observed that the best results came from representations based on Distilbert and BERTimbau with the GNB and MLP models. Fig. 6 illustrates a graph of the textual representations that performed better in each representation model of Table 5.

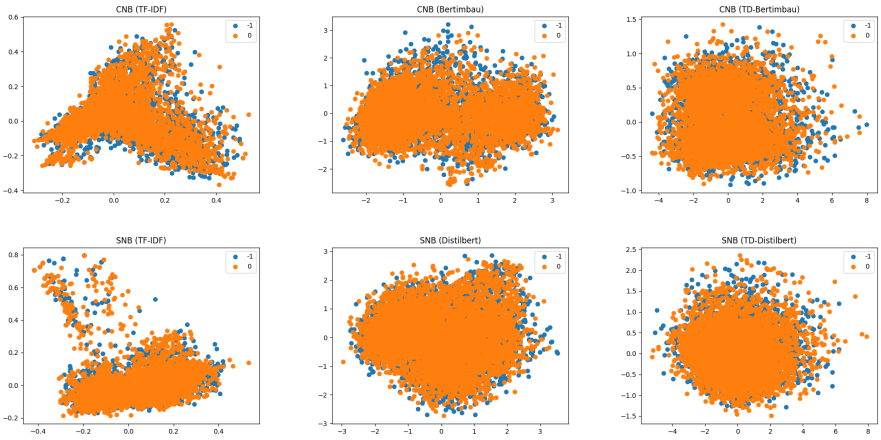

Fig. 6. CNB and SNB. PCA technique for plotting textual representations.

Confronting the investigated strategies, we found that using BoW may reduce performance. On the other hand, semantic features allow satisfactory results when considering an extensive training set. Thus, we compared only the performances of the BERT representations (BERT and Dist. B.Bau) and the proposed TD-BERT model (TD-Be, TD-Di and TD-Bau) regarding the PB and NB assessments. We can observe in Table 3 that among the best results, 75% are from BERT models and 25% are from TD-BERT models. Concerning the performances of Table 5, there was a tie of 50% for each representation model.

5 Conclusion

We introduced an automatic text labeling approach using information extracted from time series. This paper innovates by considering a weak supervision technique to label a large volume of texts. Text documents and agribusiness commodity price series were employed to assess performance in four classification scenarios (CPB, SPB, CNB and SNB). Our experimental evaluation considered nine textual representations and different learning paradigms. In addition, we proposed a text representation model that measures the distance between Terms and Documents from pre-trained BERT models (TD-BERT).

Regarding the best results of the Positive Binary and Negative Binary assessment scenarios, ten between sixteen are representations from the BERT models (62.5%). The proposed TD-BERT models performed better in some cases by analyzing neural language-based representation models. In general, neural language-based representation models outperformed BoW-based models. However, a limitation of the TD-BERT models is processing time, and future work can be conducted to reduce computational costs.

The designed labeling function can be an alternative to annotating a large volume of text documents. Automatic labeling can be imprecise but useful when many texts are not labeled. In this study, the limitation of the weak supervision analysis consisted of the class imbalance. Future research can developed strategies to propagate labels through semi-supervised learning to reinforce labeling. In addition, through connectionist approaches, other external factors can be used in the labeling function; *e.g.*, we can consider a weighting coefficient in the labeling of news.

Acknowledgements: This work was carried out at the Center for Artificial Intelligence (C4AI-USP) and partially supported by the São Paulo Research Foundation (FAPESP) (grant #2019/07665-4) and the IBM Corporation. The authors of this paper thank FAPESP (Process 2019 / 25010-5) and the National Center for Scientific and Technological Development (CNPq) (process 309575/2021-4). The corresponding author thanks the Minas Gerais State Research Support Foundation (FAPEMIG) (Process PCRH BPG-00054-210).

References

1. Aggarwal, C.C.: Machine learning for text, vol. 848. Springer (2018)
2. Alfonseca, E., Filippova, K., Delort, J.Y., Garrido, G.: Pattern learning for relation extraction with a hierarchical topic model. In: Annual Meeting of the Association for Computational Linguistics. vol. 2, pp. 54–59 (2012)
3. Anklin, V., Pati, P., Jaume, G., Bozorgtabar, B., Foncubierta-Rodriguez, A., Thiran, J.P., Sibony, M., Gabrani, M., Goksel, O.: Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 636–646. Springer (2021)
4. Araujo, A.F., Gôlo, M.P., Marcacini, R.M.: Opinion mining for app reviews: an analysis of textual representation and predictive models. Automated Software Engineering **29**(1), 1–30 (2022)

5. Batista-Navarro, R., Hawkins, O.: Topic modelling vs distant supervision: A comparative evaluation based on the classification of parliamentary enquiries. In: International Conference on Theory and Practice of Digital Libraries. pp. 415–419. Springer (2019)
6. Boecking, B., Neiswanger, W., Xing, E., Dubrawski, A.: Interactive weak supervision: Learning useful heuristics for data labeling. arXiv preprint arXiv:2012.06046 (2020)
7. Chatfield, C., Xing, H.: The analysis of time series: an introduction with R. CRC press (2019)
8. Chen, L.M., Xiu, B.X., Ding, Z.Y.: Multiple weak supervision for short text classification. Applied Intelligence pp. 1–16 (2022)
9. Dai, E., Shu, K., Sun, Y., Wang, S.: Labeled data generation with inexact supervision. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 218–226 (2021)
10. De Sa, C., Ratner, A., Ré, C., Shin, J., Wang, F., Wu, S., Zhang, C.: Deepdive: Declarative knowledge base construction. ACM SIGMOD Record **45**(1), 60–67 (2016)
11. dos Santos, B.N., Marcacini, R.M., Rezende, S.O.: Multi-domain aspect extraction using bidirectional encoder representations from transformers. IEEE Access **9**, 91604–91613 (2021)
12. Helmstetter, S., Paulheim, H.: Collecting a large scale dataset for classifying fake news tweets using weak supervision. Future Internet **13**(5), 114 (2021)
13. Hsieh, C.Y., Lin, W.I., Xu, M., Niu, G., Lin, H.T., Sugiyama, M.: Active refinement for multi-label learning: A pseudo-label approach. arXiv preprint arXiv:2109.14676 (2021)
14. Janev, V., Pujić, D., Jelić, M., Vidal, M.E.: Survey on big data applications. In: Knowledge Graphs and Big Data Processing, pp. 149–164. Springer, Cham (2020)
15. Krause, S., Li, H., Uszkoreit, H., Xu, F.: Large-scale learning of relation-extraction rules with distant supervision from the web. In: International Semantic Web Conference. pp. 263–278. Springer (2012)
16. Lison, P., Hubin, A., Barnes, J., Touileb, S.: Named entity recognition without labelled data: A weak supervision approach. arXiv preprint arXiv:2004.14723 (2020)
17. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011 (2009)
18. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid training data creation with weak supervision. In: International Conference on Very Large Data Bases. vol. 11, p. 269. NIH Public Access (2017)
19. Rekatsinas, T., Chu, X., Ilyas, I.F., Ré, C.: Holoclean: Holistic data repairs with probabilistic inference. arXiv preprint arXiv:1702.00820 (2017)
20. Shi, Y., Xiao, Y., Niu, L.: A brief survey of relation extraction based on distant supervision. In: International Conference on Computational Science. pp. 293–303. Springer (2019)
21. Shu, K., Zheng, G., Li, Y., Mukherjee, S., Awadallah, A.H., Ruston, S., Liu, H.: Leveraging multi-source weak social supervision for early detection of fake news. arXiv preprint arXiv:2004.01732 (2020)
22. Souza, F., Nogueira, R., Lotufo, R.: Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649 (2019)
23. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Brazilian Conference on Intelligent Systems (2020)

24. de Souza, M.C., Nogueira, B.M., Rossi, R.G., Marcacini, R.M., dos Santos, B.N., Rezende, S.O.: A network-based positive and unlabeled learning approach for fake news detection. Machine Learning pp. 1–44 (2021)
25. Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q., Gao, J.: Weak supervision for fake news detection via reinforcement learning. In: AAAI Conference on Artificial Intelligence. vol. 34, pp. 516–523 (2020)
26. Yao, W., Liu, J., Cai, Z.: Personal attributes extraction in chinese text based on distant-supervision and lstm. In: Advances in Computer Science and Ubiquitous Computing, pp. 511–515. Springer (2017)
27. Zhou, Z.H.: A brief introduction to weakly supervised learning. National Science Review **5**(1), 44–53 (2018)