Classifying Emotions in Tweets from the Financial Market: A BERT-based Approach

Wesley Pompeu de Carvalho
School of Arts, Sciences and Humanities
University of São Paulo
São Paulo, Brazil
wesley.carcalho@usp.br

Norton Trevisan Roman
School of Arts, Sciences and Humanities
University of São Paulo
São Paulo, Brazil
norton@usp.br

Abstract

Behavioural finance emphasizes the relevance of investor sentiment and emotions in the pricing of financial assets. However, little research has examined how discrete emotions can be detected in text related to this domain, with extant work focusing mostly in sentiment instead. This study approaches this problem by describing a framework for emotion classification in tweets related to the stock market written in Brazilian Portuguese. Emotion classifiers were developed, based on Plutchik's psychoevolutionary theory, by fine-tuning BERTimbau, a pre-trained BERT-based language model for Brazilian Portuguese, and applying it to an existing corpus of tweets from the stock market domain, previously annotated with emotions. The results demonstrated statistically significant improvements over a stratified random baseline (Welch's *t*-test, $p \ll 0.001$ across all axes), with macro-F1 scores ranging from 0.50 to 0.61. These findings point to the feasibility of using transformer-based models to capture emotional nuance in financial texts written in Portuguese and provide a reproducible framework for future research.

1 Introduction

During the 1980s and 1990s, Behavioural Finance (Illiashenko, 2017) gained prominence by highlighting the role of economic and behavioural factors in financial decision-making and the formation of asset prices in financial markets. In recent decades, investor sentiment has been recognized as a major driver of market dynamics (Long et al., 1990; Mao et al., 2011), motivating the need for reliable and scalable methods to assess its impact.

With the rise of social media over the past two decades, large volumes of textual data became available online, encouraging the development of techniques to measure investor sentiment and emotions. In this context, the field has evolved into two main areas: sentiment analysis, which seeks to identify the sentiment – usually positive, negative, or neutral – expressed by a specific audience toward a topic or object of interest (Albu and Spînu, 2022; Pang and Lee, 2008); and emotion analysis, which identifies specific emotions conveyed by authors (da Silva et al., 2020), usually under some psychological framework.

Although sentiment analysis has become widespread in the financial domain, fewer studies have focused on emotion analysis. The reason for this behaviour might lie perhaps in the fact that there is no universally accepted model of emotion. In this case, the existence of competing models, such as Ekman's (Ekman, 1992) and Plutchik's Wheel of Emotions (Plutchik and Kellerman, 2013) for example, contributes to the scarceness of annotated corpora.

Within this context, one such corpus stands out by presenting a set of tweets from the stock market domain, written in Brazilian Portuguese and manually annotated with emotions according to the Plutchik's Wheel of Emotions (da Silva et al., 2020). This corpus was later on manually curated (whereby duplicates and tweets not related to the stock market domain were removed) and augmented with Part-of-Speech information (Felippo et al., 2023).

In our work, we started out from these corpora, and developed a set of four automatic emotion classifiers for tweets based on BERTimbau (Souza et al., 2023) – a pre-trained BERT-based language model for Brazilian Portuguese. Within our framework, each of Plutchik's four main emotional axes was modelled as a ternary classification problem. For each axis, 30 independent training iterations were executed using a repeated holdout strategy with different train/test splits in each iteration. In every iteration, hyperparameter tuning was performed via 10-fold stratified cross-validation on

Proceedings of Recent Advances in Natural Language Processing,pages 218–226
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-026

the training set to identify the best configuration. A final model was then retrained using the selected hyperparameters and evaluated on a hold-out test set, generating a distribution of macro-F1 scores in out-of-sample data.

In following this procedure, our contribution to the field is twofold. First, we introduce an automatic emotion classifier for stock market tweets written in Brazilian Portuguese, along with a reproducible training and validation framework. Second, we present a detailed evaluation of the classifier's performance on the target corpus, including interval estimates and a robust statistical analysis to assess its expected performance on similar datasets.

The rest of this article is organised as follows. Section 2 presents some related work, for both sentiment and emotion classification, also giving a historical overview on the subject. Section 3, in turn, summarises the fundaments of Plutchik's theory. Our study's experimental setup is described in Section 4, along with the step-by-step procedure we followed to train the system and assess its performance. Obtained results are described and discussed in Section 5. Finally, in Section 6 we present our final remarks and directions for future work.

2 Related Work

In the early 2000s, researchers in sentiment and emotion analysis relied primarily on lexical dictionaries to classify texts from news articles, microblogs, company documents, and other sources (*e.g.* Frank and Antweiler (2001)). Over time, simpler Natural Language Processing (NLP) techniques, such as bag-of-words and n-grams, became common approaches for extracting numerical representations from textual data (*e.g.* Mao et al. (2011)).

Along with these lexical approaches, users' behaviour, especially in microblogging texts, also allowed for the use of non-lexical resources, such as emojis for example, as in Davidov et al. (2010). In this case, the authors explored automatic sentiment labelling techniques using hashtags and emojis to assign sentiments to messages, resulting in a sentiment classifier that achieved F1-scores of 0.80 and 0.86 in binary classification tasks, outperforming the random baseline.

From 2013 onwards, with further progress in NLP and Machine Learning techniques, the use of distributed representations such as *word embeddings* became widespread. These embeddings represent texts as continuous vectors built through deep learning techniques (Almeida and Xexéo, 2023), and were applied in research such as the one reported by Pagolu et al. (2016), where a sentiment classifier for tweets was trained based on *Word2Vec* vectors, achieving a 70% accuracy rate in the sentiment classification task.

More recently, pre-trained language models (PLMs) based on the Transformer architecture have achieved state-of-the-art performance across a range of NLP tasks (Zhao et al., 2024). One such model, BERT (Devlin et al., 2019), released in 2018, became widely used due to its adaptability to various tasks across specific domains, including sentiment analysis in the financial domain. An example of applying this model in finance can be found in Dong et al. (2020), where BERT was fine-tuned for the task of sentiment analysis on stock market-related tweets, achieving an accuracy of 84.5%. Another example is FinBERT (Araci, 2019), a model specifically adapted to the financial domain through further pre-training BERT on a financial corpus, which achieved an accuracy of 86% and a macro-F1 score of 0.84 on the Financial PhraseBank dataset.

Regarding the automatic classification of emotions, although operating on audio data rather than text, one finds the study by Hajek and Munk (2023), who built an emotion classifier based on a convolutional neural network (CNN) trained on the RAVDESS dataset (Livingstone and Russo, 2018) – an audio dataset in English created for general affective computing research. Using this dataset, the CNN achieved an accuracy of 69.8% on the test set through fivefold cross-validation. The trained model was later applied to earnings conference calls to extract managerial emotions, which were then used as features in a financial distress prediction framework.

Another relevant contribution is presented by da Silva (2020), who developed an emotion classifier for tweets in Brazilian Portuguese, focusing on the financial domain. His approach involved training a support vector machine (SVM) with syntactic and lexical features on a general corpus annotated via distant supervision and testing it on the manually annotated corpus of tweets presented in da Silva et al. (2020). The presented model outperformed traditional baselines in five out of the eight emotions analysed.

3 Plutchik's Wheel of Emotions

Emotion is a multifaceted concept that is hard to delineate, with definitions ranging from responses to delayed or inhibited actions (Lang, 1995) to bodily changes triggered by the brain when interpreting external scenarios (Damásio, 1994). Not surprisingly, so lacks emotional classification a universally accepted model, with many competing theories and taxonomies being at use. Widely used frameworks include Ekman's set of six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) (Ekman, 1992) and Plutchik's multidimensional model (Plutchik and Kellerman, 2013; Cowie and Cornelius, 2003), which defines emotions along a continuous space defined by four opposing basic axes: *Joy-Sadness*, *Anger-Fear*, *Trust-Disgust*, and *Anticipation-Surprise*.

Neighbouring axes can then be combined to form new axes (secondary, tertiary and so on) and, consequently, to define new emotions in the continuous space. This naturally leads to an infinite set of possibilities, making this theory hard (if not impossible, in its full extent) to model in a computational way. This is probably the reason why practical studies often focus on a simplified subset (Graterol et al., 2021) of Plutchik's emotions. In our case, and following (da Silva, 2020), we focus on emotions belonging to the four basic axes. This theoretical model was chosen because it includes surprise and anticipation, emotions seen as crucial in finance, where stock prices react to forecasts or unexpected events (da Silva, 2020). The idea of opposing emotions also guided the annotation process and supports future correlation studies (Plutchik and Kellerman, 2013).

4 Experimental Setup

Two versions of the same corpus of financial market tweets written in Brazilian Portuguese build our materials. The first version comprises a set of tweets related to the Brazilian stock market, manually labelled according to Plutchik's Wheel of Emotions (da Silva et al., 2020). The corpus comprises 4,277 tweets[1], automatically collected between March and May 2014, containing codes for at least one of the 73 stocks comprising the Ibovespa index at the time. It was then manually annotated by at least three different volunteers, with each tweet being annotated four times – once for each emotional axis (*trust* × *disgust*, *joy* × *sadness*, *anticipation* × *surprise*, and *anger* × *fear*).

This corpus was later revisited in another research, aiming to augment it with morphosyntactic information in a stand-off manner (*cf.* Felippo et al. (2023)). During this effort, some duplicated tweets were found, along with other tweets that were not related to the stock market. These were removed, resulting in the 3,994 tweets of the DANTEStocks corpus. In our research, we started from DANTEStocks to obtain the textual content of the tweets, and then followed its links to the corpus in da Silva et al. (2020) to fetch the emotions associated to each of DANTEStocks' tweets.

With the corpus at hand, we turned our attention to building and evaluating an automatic system for emotion classification. To this end, we decided to use BERTimbau (Souza et al., 2023) – a version of BERT pre-trained in Brazilian Portuguese texts. This model was chosen not only for its applicability to many NLP related tasks in Portuguese, but also to allow for the experiment's reproducibility, by downloading the model and keeping it fixed. This control over a possible confounding factor (in this case, the model's configurations) is something that cannot be guaranteed in API-based alternatives, for example. Hence, we took BERTimbau's cased version[2] and fine-tuned it for emotion classification in this corpus, following the steps described in Procedure 1.

The main idea in following this procedure was to get a distribution of macro-F1 values at different test sets, so as to have an idea of the classifiers' variance. This was done by fine-tuning the PLM in a training set and measuring its performance at the test set, also exploring best hyperparameter combinations for each emotional axis. These combinations were determined through 10-fold cross validation at the training set. This whole procedure was then repeated 30 times, to get the desired distribution.

The procedure then begins by loading the pretrained language model and setting the hyperparameters (Table 1) to be explored during the validation phase (steps 2 and 3). Since our intent was to train a different classifier for each emotional axis, as done in da Silva (2020), the next step was to define the list of possible axes (step 4).

[1] Freely available at https://www.kaggle.com/datasets/fernandojvdasilva/stock-tweets-ptbr-emotions

[2] https://huggingface.co/neuralmind/bert-base-portuguese-cased

Procedure 1 Classifier's Training and Evaluation

```
1:  procedure ModelTraining
2:      $Mod\_pre$ ← pretrained language model
3:      $Grid\_hiperp$ ← Range of hyperparameter values
4:      $Emotion\_axis$ ← Plutchik's emotional axes
5:      $Seed\_init$ ← Initial seed for the random number generator
6:      $N\_rep$ ← 30 (Number of times the repeated holdout process is executed)
7:      $Seeds$ ← Array with N_rep (pseudo) random seeds, generated from Seed_init
8:      for $i \leftarrow 1$ to $N\_rep$ do                                    ▷ Repeated Holdout
9:          $Seed_i$ ← Seeds[i]
10:         $(Dtr, Dts)$ ← Data split into training (80%) and test (20%) sets using $Seed_i$
11:         $Seed_{kf}$ ← Seed for 10-fold, ensuring the same folds
12:         for axis $e$ in $Emotion\_axis$ do
13:             for hyperparameter $hp$ in $Grid\_hiperp$ do
14:                 stratified 10-fold cross-validation in $Dtr$ using $Seed_{kf}$ as seed
15:                     Strategy: Undersampling of the training folds.
16:                     $MF1_{k\_hp\_e\_i}$ ← Best macro-F1 for $hp$, fold $k$, and axis $e$
17:                 $MMF1_{hp\_e\_i}$ ← Arithmetic mean across the $k$ values of $MF1_{k\_hp\_e\_i}$
18:             $hp_{e\_i}$ ← Best $hp$ for $e$                   ▷ Criterion: highest $MMF1_{hp\_e\_i}$
19:             $Mod\_Fin\_hp_{e\_i}$ ← Model retrained with $Dtr$ using $hp_{e\_i}$
20:                 Strategy: Undersampling of the $Dtr$ set
21:             $Aval_{e\_i}$ ← Array with the performance of $Mod\_Fin\_hp_{e\_i}$ on $Dts$ at each epoch
```

Table 1: Hyperparameter values tested during validation

Hyperparameter	Values
initial_learning_rate	$[5 \times 10^{-5}, 5 \times 10^{-6}]$
warmup_steps	[100, 500]
train_batch_size	[8, 32]

In the sequence, and as a way to try to reduce the influence of any bias introduced by the arbitrary definition of random seeds in the system, while still allowing for the experiment's reproducibility, we set an initial seed (step 5), which was then used to generate other 30 seeds[3]. Then, at each of the 30 repetitions of the holdout procedure, a different seed is used to create distinct train/test splits (steps 6 to 10), with 80% of the data being reserved for training and the remaining 20% building the test set.

In order to have cross-validation run on the same folds for all classifiers, another seed was fixed (step 11). Then, for each of the four emotional axes under consideration, we built a different classifier for each parameter combination, based on the definitions and values established in the hyperparameter grid (steps 12 to 21), and modelling each axis as a three-class classification problem (Emotion$_1$, Emotion$_2$, and Neutral).

For each hyperparameter combination, 10-fold cross-validation is performed (step 14) in its corresponding classifier using the training set[4] defined in step 10 and the cross-validation seed set in step 11. The cross-validation strategy is executed as follows:

1. Folds are created by stratified random sampling;

2. Another version of the training set is then built by *Undersampling* the majority class in the training folds, while keeping the stratified distribution of the validation fold unchanged; and

3. The classifier is fine-tuned in the undersampled training folds and evaluated on the validation fold. Training stops early if validation macro-F1 does not improve for three consecutive epochs, or after a maximum of 20 epochs

[3] This number was chosen due to the long processing time spent in each iteration

[4] *I.e.* the training set is further split into training$_2$ and validation sets

is reached.

For each hyperparameter configuration, the mean macro-F1 is computed across the best validation scores in each fold. At the end of the cross-validation procedure, the optimal hyperparameter setting, for each specific emotional axis, is selected based on the highest average validation performance (steps 17 and 18).

Using these selected hyperparameters, a new instance of the classifier is then trained on the entire training set (D_{tr}) after undersampling it (step 22). Training is stopped if the training macro-F1 does not improve for two consecutive epochs or after 20 epochs are run. This stopping criterion was followed to reduce the risk of underfitting, even at the price of increasing the risk of overfitting which, in turn, can be verified from the evaluation of the model at the the independent test set (step 21).

Finally, as a benchmark for comparison, along with the main classifiers four random classifiers were also built, one for each of the main emotional axes. Each classifier randomly assigned labels according to their observed frequency in the training data[5].

5 Results and Discussion

The distribution of mean macro-F1 scores across the 30 repetitions of the experiment are shown in Figure 1. In this figure, the line in black shows the theoretical Normal curve for the data (*i.e.* the curve obtained with the distribution's mean and standard deviation), whereas the light gray curve illustrates a kernel density estimate, providing an approximate view of the shape of the underlying distribution.

The differences between our main classifiers and their random counterparts, for each axis, can be seen in Table 2, where μMF1 denotes the mean Macro-F1 score and σ represents the standard deviation across all repetitions. These differences were found to be significant, as shown in an independent samples Welch's t-test[6], whose results for each emotional axis can be seen in Table 3.

Welch's test was chosen for not assuming variance homogeneity, something that was rejected with a Levene's test, adjusted for multiple testing at the 95% confidence level. The test was conducted unilaterally, considering only the alternative hypothesis that the developed models perform better than the random classifier. The statistical decision was made with an adjusted significance level of $\alpha = 1.25\%$, resulting from the Bonferroni correction for multiple testing. Finally, Table 4 presents a confidence interval for mean macro-F1 values at each emotional axis. The interval was estimated with a 95% confidence level from a Student's t-distribution.

Table 2: Comparison of mean Macro F1 Scores in test between Developed and Random Classifiers

Axis	Classifier	μMF1	σ
ANG-FEA	Developed	**0.52**	0.03
	Random	0.35	0.01
JOY-SAD	Developed	**0.59**	0.02
	Random	0.33	0.01
TRU-DIS	Developed	**0.61**	0.02
	Random	0.32	0.01
ANT-SUR	Developed	**0.50**	0.02
	Random	0.33	0.01

Table 3: Welch's t-test for the difference between our classifiers and their random counterparts.

Axis	T	p	Result
ANF-FEA	30.7	$\ll 0.1\%$	H_0 rejected
ANT-SUR	37.8	$\ll 0.1\%$	H_0 rejected
JOY-SAD	52.4	$\ll 0.1\%$	H_0 rejected
TRU-DIS	73.3	$\ll 0.1\%$	H_0 rejected

Our results illustrate the differences amongst the emotional axes, with *TRU-DIS* and *JOY-SAD* being over 0.59 macro-F1 and *ANF-FEA* and *ANT-SUR* being under 0.52. A breakdown analysis of the F1-score distribution for each class, within each axis, across the 30 experiments (Figure 2), revealed that, for *JOY-SAD* and *ANG-FEA*, the *Neutral* class performed considerably better than the other classes, thereby raising their *macro-F1* average. This effect is also present, although less pronounced, in *ANT-*

Table 4: Means and confidence intervals for Macro F1 in test data

Axis	Mean	CI (95%)
ANF-FEA	0.52	[0.51, 0.53]
JOY-SAD	0.59	[0.58, 0.59]
TRU-DIS	0.61	[0.60, 0.62]
ANT-SUR	0.50	[0.50, 0.51]

[5] Both random and main classifiers were trained and tested in the same data sets.

[6] The normality of the F1-score distribution for both classifiers was verified using the Shapiro-Wilk test. In none of the cases was the null hypothesis rejected at a 5% confidence level, with Bonferroni adjusting for multiple testing.

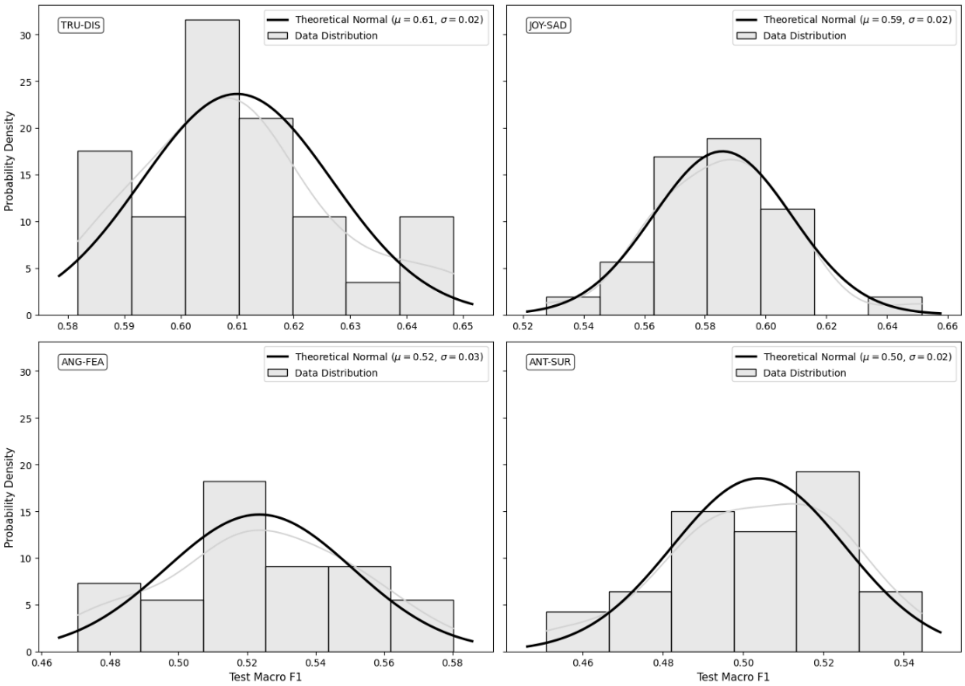

Figure 1: Distribution of macro-F1 scores at the test set for Trust × Disgust, Joy × Sadness, Anger × Fear and Anticipation × Surprise

SUR. Conversely, *TRU-DIS* – our highest macro-F1 score – exhibited greater balance among classes. These results evidence the practical difficulty in classifying emotions in this corpus, with some axes standing out of others.

However modest the obtained results, in all axes our model outperforms its random counterpart, as already shown in Table 2, being from 49% (*ANG-FEAR*) to 91% (*TRU-DIS*) higher in macro-F1 than this baseline, a difference found to be significant (as already pointed out in Table 3). A possible reason for this outcome lies in the inherent complexity of the task, which requires classification into three distinct classes while simultaneously dealing with texts from social media.

Another point to be considered is that classes are unbalanced in the test set (recall that test sets were not balanced), as shown in Figure 3. Interestingly, the most unbalanced axes in Figure 3 are those in which *neutral* (*i.e.* the majoritarian class) was found to perform better, as shown in Figure 2. This is an indicative of the importance of class imbalance in our results.

In this sense, even though the balancing of the classes at the training set led the model to better learn patterns in the minoritarian classes, as evidenced by mean macro-F1 values at training being higher than their test counterparts (Table 5), the difference in distribution between training and testing sets might have played a role in these results. Although results might have been better should we balance the test sets, we decided to keep them close to the original distribution so that the obtained estimates at these sets are kept unbiased.

Table 5: Comparison of Train and Test Macro-F1 Scores for Each Emotional Axis

Axis	Split	μMF1	σ
ANG-FEA	Train	0.64	0.04
	Test	0.52	0.03
JOY-SAD	Train	0.74	0.03
	Test	0.59	0.02
TRU-DIS	Train	**0.83**	0.02
	Test	**0.61**	0.02
ANT-SUR	Train	0.74	0.02
	Test	0.50	0.02

Additionally, the limitation in the exploration of the hyperparameter space may have contributed to these results. The decision to vary only a few parameters (for example, batch size, learning rate, and number of epochs) was mainly motivated by the computational costs involved in building, validating, and testing multiple instances of the model, especially under the existing hardware capacity constraints. It is also noteworthy that the choice for

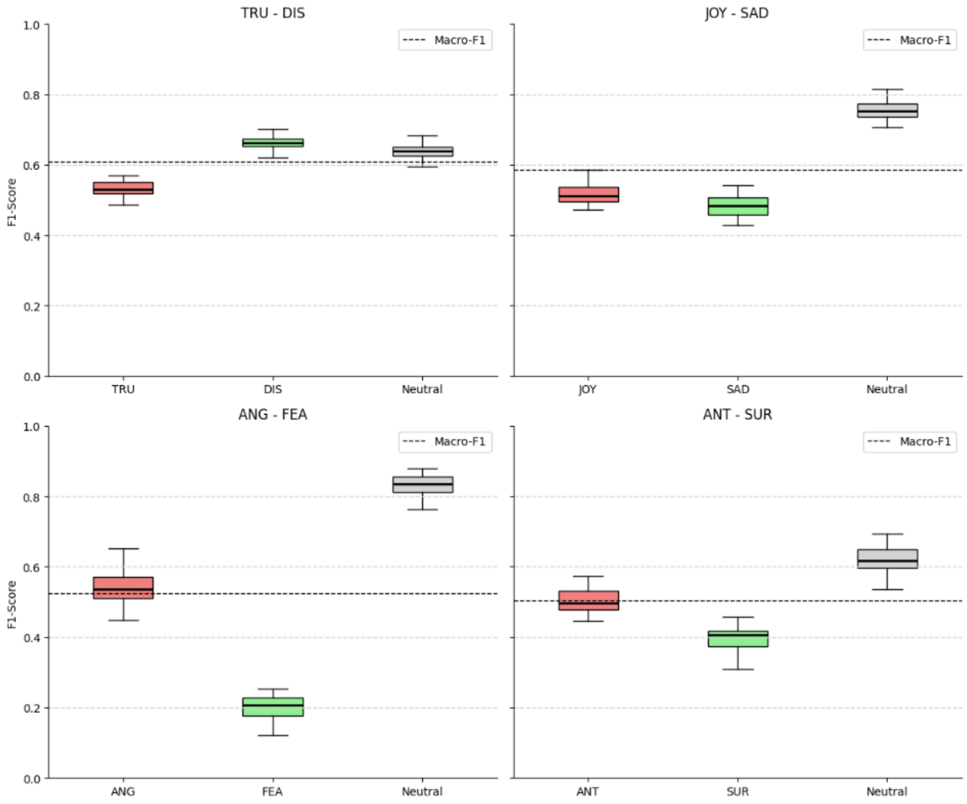

Figure 2: Breakdown of F1-scores for each emotional axis. The dashed line represents the Macro-F1 Score for the axis.

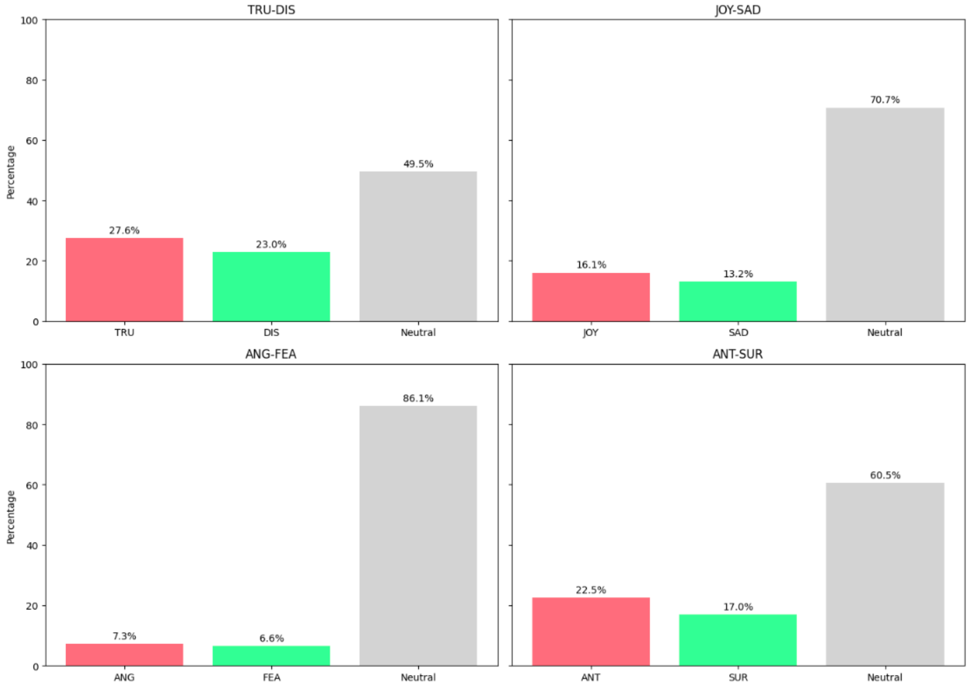

Figure 3: Test data distribution for the emotional axes.

specific values for these hyperparameters was, in part, arbitrary, indicating that a broader variation of configurations could potentially lead to superior performance. Thus, exploring other hyperparameters, as well as expanding the range of considered values, emerges as a relevant prospect for future work.

Finally, the same hardware constraints that limited our choice for hyperparameters also played an important role in our decision to go for BERTimbau's base version instead of a larger language model. In this sense, experiments with this model's large version, or any other language model, also comes out as an interesting venue for future improvement, so as to verify whether increasing the encoder's capacity could result in significant improvements in classifier performance.

6 Conclusion

In this article we described our efforts in developing and evaluating an automatic emotion classifier for stock market tweets written in Brazilian Portuguese, under the Plutchik's wheel of emotions paradigm, furnishing a detailed analysis of out-of-sample performance. Our results show that the adopted framework enabled a systematic and in-depth evaluation of classifier performance in each of Plutchik's emotional axes.

The combined use of *repeated holdout* and cross-validation ensured both the variability of training and testing partitions and the consistent selection of best hyperparameters, mitigating the risk of overfitting to a single data split. Furthermore, by enabling the execution of multiple experiments, the framework provided greater statistical robustness, allowing for the assessment of whether the classifiers were able to extract knowledge from the data.

A comparison of the developed models against random benchmarks showed that, although our models outperformed the baseline, performances varied across the different emotional axes, with *TRU-DIS* and *JOY-SAD* presenting the best performances, while *ANF-FEA* and *ANT-SUR* proved more challenging. Moreover, the observed trend that had the *Neutral* class to be classified with greater confidence reinforces the need for strategies to improve the differentiation between emotion pairs.

Among the limitations of the study, the restricted exploration of the hyperparameter space stands out, conditioned by computational costs. Future experiments may consider a broader search for optimized hyperparameters, in addition to testing more complex architectures, such as the *large* version of BERTimbau.

The study then demonstrates the feasibility of the proposed approach for emotion analysis and highlights challenges to be overcome to enhance classifier performance. The results obtained serve as a basis for future investigations, pointing to paths for the development of more robust and adaptable models for this type of task.

Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Ionuț-Alexandru Albu and Stelian Spînu. 2022. Emotion detection from tweets using a bert and svm ensemble model. *arXiv preprint arXiv:2208.04547*.

Felipe Almeida and Geraldo Xexéo. 2023. Word embeddings: A survey.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Roddy Cowie and Randolph R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1):5–32.

António Damásio. 1994. *DESCARTES' ERROR: emotion, reason, and the human brain*. Putnam.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249, Beijing, China. Coling 2010 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yingzhe Dong, Da Yan, Abdullateef Ibrahim Almudaifer, Sibo Yan, Zhe Jiang, and Yang Zhou. 2020. Belt: A pipeline for stock price prediction using news.

In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1137–1146.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Ariani di Felippo, Norton Trevisan Roman, Thiago Alexandre Salgueiro Pardo, and Lucas Panta de Moura. 2023. The dantestocks corpus: an analysis of the distribution of universal dependencies-based part-of-speech tags. *Revista da ABRALIN*, 22(2):249–271.

Murray Z. Frank and Werner Antweiler. 2001. Is all that talk just noise? the information content of internet stock message boards. AFA 2002 Atlanta Meetings, Sauder School of Business Working Paper.

Wilfredo Graterol, Jose Diaz-Amado, Yudith Cardinale, Irvin Dongo, Edmundo Lopes-Silva, and Cleia Santos-Libarino. 2021. Emotion detection for social robots based on nlp transformers and an emotion ontology. *Sensors*, 21(4):1322.

Petr Hajek and Michal Munk. 2023. Speech emotion recognition and text sentiment analysis for financial distress prediction. *Neural Computing and Applications*, 35(29):21463–21477.

Pavlo Illiashenko. 2017. Behavioral finance: History and foundations. *Visnyk of the National Bank of Ukraine*, (239):28–54.

Peter J. Lang. 1995. The emotion probe: Studies of motivation and attention. *American psychologist*, 50(5):372.

Steven R. Livingstone and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess). *PloS one*, 13(5):e0196391.

J. Bradford De Long, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. 1990. Noise trader risk in financial markets. *Journal of Political Economy*, 98(4):703–738.

Huina Mao, Scott Counts, and Johan Bollen. 2011. Predicting financial markets: Comparing survey, news, twitter and search engine data.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Robert Plutchik and Henry Kellerman. 2013. *Theories of Emotion*, volume 1. Academic Press.

Fernando José Vieira da Silva. 2020. *Cross-Domain Emotion Detection in Tweets*. Ph.D. thesis, Universidade Estadual de Campinas.

Fernando José Vieira da Silva, Norton Trevisan Roman, and Ariadne M.B.R. Carvalho. 2020. Stock market tweets annotated with emotions. *Corpora*, 20(3):343–354.

F.C. Souza, R.F. Nogueira, and R.A. Lotufo. 2023. Bert models for brazilian portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, 149:110901.

Wayne Xin Zhao et al. 2024. A survey of large language models.