

Exploring financial sentiment analysis with the Financial Phrasebank dataset

Hugo Veríssimo

Complements of Machine Learning 24/25
University of Aveiro
Aveiro, Portugal
hugoverissimo@ua.pt

João Cardoso

Complements of Machine Learning 24/25
University of Aveiro
Aveiro, Portugal
joaopcardoso@ua.pt

Abstract—abstratootototot

Keywords: key, word, number, 1

I. INTRODUCTION

With the ever increasing volume of information created and distributed by the minute, it is more important than ever to have access to fast and reliable analysis of any information we may come across. Especially with the democratized access to financial instruments and capital markets, where individuals have the possibility to invest in virtually any company on the stock exchange, it is important to have ways to leverage against giant institutions with hundreds of financial analysts at their disposal.



Fig. 1: Power to the people, colorized (circa 1917).

Historically, financial analysis relied heavily on fundamental analysis (examining earnings, balance sheets, annual financial reports), which required extensive knowledge in the field (also the strategy that made Warren Buffett one of the richest men in the world), along with technical analysis (studying price and volume trends). Around 2010, after the 2008 global financial crisis, there was a surge in news analysis to evaluate the tone and derive investment strategies from it. Due to the lack of domain specific lexicon these analysis were falible, until the work by Loughran and McDonald was published, a financial lexicon based on 10-K forms (i.e., annual financial reports) and dictionaries. This allowed to use more sofisticated analysis rather than using the presence of negative words as a signal to sell.

Upon the launch of Twitter, information streams increased dramatically, making more and more data available for analysis. But, machine learning was not heavily used, as most

data was not annotated, or there was very little data with high-quality annotations. In 2014, P. Malo *et al.* published a fundamental dataset for financial sentimental analysis, that is still used, the Financial Phrasebank. It is unique, for the inclusion of important aspects as directional expressions (e.g., profits decreased), entity polarity shifts (e.g. profits may be negative if decreased), and phrase level context.

With this, machine learning models started finding their place in research, as the field of natural language processing (NLP) grew and niche fields such as financial investments found useful data. In this work we explore the Financial Phrasebank dataset, by implementing different machine learning and deep learning models to evaluate the sentiment of sentences related to financial news.

II. STATE OF THE ART

The field of NLP has grown drastically in the past decade, progressing from recurrent neural networks (RNN) and related models such as Long-Short Term Memory (LSTM), to the transformers-type models, large language models (LLM) and text generative models as ChatGPT. With the Financial Phrasebank the field expanded into financial analysis, with several works of relevance being published in recent years.

In the work of Araci (2019), the author developed a BERT-based model trained specifically on texts with financial data. BERT, Bidireccional Encoder Representations from Transformers, is a large language model developed by Google (2018), benefitting largely from the fact that it can "hold" in memory large chunks and in both directions, simultaneously. The fact that it is built on the Transformer encoder architecture, it can weigh the importance of different words in a sentence by using a self-attention. The model is pre-trained on large unlabelled corpora (e.g., Wikipedia, BookCorpus), and can be fine-tuned for specific purposes. In this work, the end model was trained on domain specific corpus such as TRC2-financial data and financial specific texts (over 440 000 sentences), and then fine tuned with the Financial Phrasebank. The model achieved an accuracy of 97 % on the Financial Phrasebank dataset with 100 % agreement, but 86 % and 85 % respectively on the dataset with all levels of agreement (the agreement will be further detailed in the Methodology

section). Later on the model was further improved by Sun *et al.* (2025), EnhancedFinSentiBERT, by including dictionary embeddings, expanded corpus that deversified the pre-training stage drastically, and a novel neutral sentiment module, that further enhanced the distinction between neutral and weak sentiments, resulting in improved accuracy (89 %) and F1-score (88 %). The pre-training stage benefited from the large diversity of the corpus, going from a few million tokens to 2.4 B tokens with the latest version.

In a similar direction, but at the fine tuning level, Atsiwo (2024) improved the data used in fine tuning, considering that most datasets have relatively short sentences (< 100 tokens), failing to leverage the full context window of LLMs like BERT (512 tokens). This was achieved by augmenting the training data with synthetic sentences generated by GPT-4, with accuracy of 89 % and F1-score of 88 % for 50 % agreement dataset.

III. METHODOLOGY

A. Dataset

https://huggingface.co/datasets/takala/financial_phrasebank [1], [2]

The Financial PhraseBank is a widely used benchmark dataset for financial sentiment analysis. It consists of roughly 4,840 English sentences (mostly news headlines or short statements) about companies, drawn from financial news articles and press releases. Each sentence is labeled with one of three sentiment classes – positive, negative, or neutral – representing the sentence's sentiment from the perspective of an investor.

TABLE I: data distribution

Sentiment	Agreement		
	50%	66%	75%
Negative	604	514	420
Neutral	2879	2535	2146
Positive	1363	1168	887
Total	4846	4217	3453
			All

o dataset foi anotado (labels) por 16 anotadores com background em economia/finanças, sendo que cada anotou um subset de frases. cada instância foi anotada por 5-8 anotadores independentes, levando então à criação de 5 sub datasets tendo em conta o nível de concordância entre as anotações: 50% de agreement, 66%, 75%, e concordância total. a distribuição das classes juntamente com a quantidade total de instancias é apresentada na tabela I. importa também referir que a cada determinada concordância é um mínimo, ou seja, por exemplo o dataset de 50% de Agreement contem os datasets de 66%, 75% e total concordância.

estes subsets permitem aos investigadores ter um equilíbrio entre quantidade de dados e qualide dos dados, representando um trade-off muito comum no espaço da aprendizagem automática (ig)

B. Exploratory Data Analysis

tendo em conta os vários subsets disponíveis do Financial Phrasebank, o escolhido para análise mais detalhada ao longo deste projeto foi o dataset 75Agreement, representado possivelmente o melhor equilibrio entre qualidade e quantidade, tendo em conta a qnt de instancia por classe face às restantes concordâncias.

DISTRIBUICAO DO SENTIMENTO - meter aqui plot

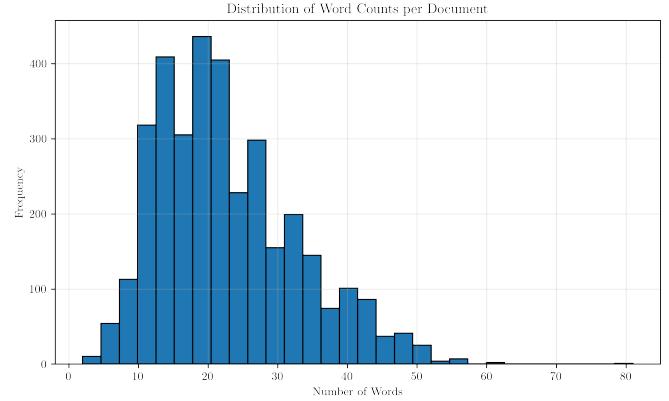


Fig. 2: word count distribution



Fig. 3: most common words in the sentences

most relevant words for class identification by checking which words appear much more in one class than others: down, decreased, profit, fell, rose

C. Preprocessing

split teste e treino

IV. MODELS

abcbdhddhd

A. model 1

model 1

B. model 2

model 2

C. model 3

model 3

V. EXPLORING BERT

A. base model

TABLE II: Hyperparameter space for

Hyperparameter	Possible Values
Epochs	{1, 2, 3, 4, 5}
Learning rate	[10^{-5} , 10^{-2}]
Weight decay	[0, 0.5]

best hyperameters:

- Num train epochs: 2
- Learning rate: 0.0001
- Weight decay: 0.1

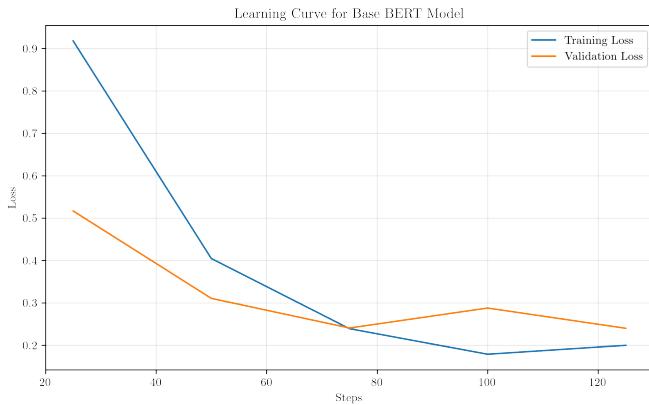


Fig. 4: learning curve

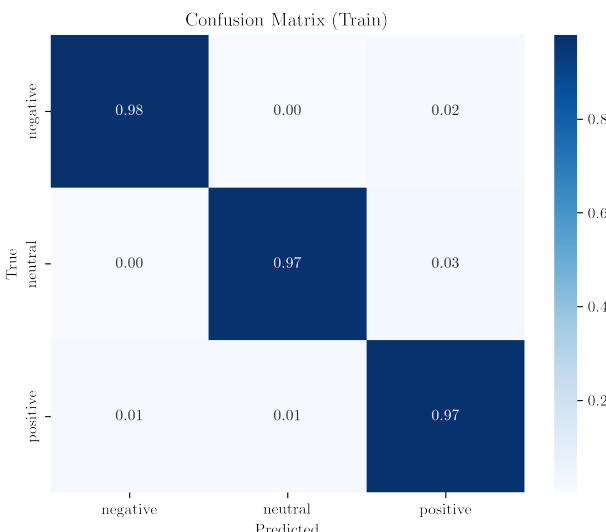


Fig. 5: base bert conf matri train

TABLE III: Classification report for BASE BERT on training data.

Class	Precision	Recall	F1-Score	Support
Negative	0.98	0.98	0.98	336
Neutral	0.98	0.97	0.98	336
Positive	0.96	0.97	0.96	336
Accuracy			0.97	1008
Macro avg	0.97	0.97	0.97	1008
Weighted avg	0.97	0.97	0.97	1008

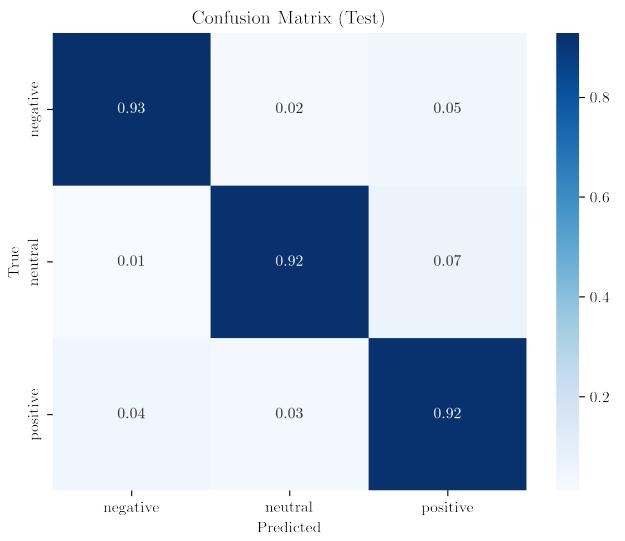


Fig. 6: base bert cong matrix test

TABLE IV: Classification report for BASE BERT on test data.

Class	Precision	Recall	F1-Score	Support
Negative	0.86	0.93	0.89	84
Neutral	0.98	0.92	0.95	429
Positive	0.84	0.92	0.88	178
Accuracy			0.92	691
Macro avg	0.89	0.92	0.91	691
Weighted avg	0.93	0.92	0.92	691

VI. DISCUSSION

results discuttion
eemplo de comparacao

TABLE V: Error metric (MAE) for the fine-tuned models, along with the best performers in the competition.

Model	MAE (Test Set)
Hybrid	0.8434
Obj. Det.	1.2645
Inst. Seg.	1.3415
Team Lacuna (1st)	0.3299
K_Junior (2nd)	0.5698

VII. CONCLUSION
conclisao



Fig. 7: Enter Caption

WORK LOAD

Both authors contributed equally to the project.

REFERENCES

- [1] P. Malo, A. Sinha, P. Takala, P. Korhonen, and J. Wallenius, “Financialphrasebank-v1.0,” 07 2013.
- [2] ——, “Good debt or bad debt: Detecting semantic orientations in economic texts,” 2013. [Online]. Available: <https://arxiv.org/abs/1307.5336>