Small Language Models for Emotion Recognition in Polish Stock Market Investor Opinions

Bartłomiej Koptyra, Marcin Oleksy, Ewa Dzięcioł, and Jan Kocoń
Department of Artificial Intelligence, Wroclaw Tech, Poland
{bartlomiej.koptyra, marcin.oleksy, ewa.dzieciol, jan.kocon}@pwr.edu.pl

***Abstract*—In this paper, we explore the application of small language models for emotion recognition in Polish stock market investor opinions. Emotion recognition has been shown to enhance stock price prediction models by providing meaningful text features. We utilize publicly available pre-trained transformer models and fine-tune them for emotion classification in Polish business articles related to WIG20, the Polish equivalent of S&P 500. Given the scarcity of domain-specific pre-trained models for Polish, we experiment with different transformer architectures, comparing their performance in recognizing emotions such as anger, anticipation, joy, sadness, and trust. Our findings indicate that the choice of a pre-trained model significantly affects performance, with the Polish RoBERTa model yielding the best results for both sentence and document-level emotion classification. We also discuss the challenges of class imbalance and the potential for improving results through additional pre-training on domain-specific data. This work contributes to developing emotion classification models for financial text in Polish and improving stock market prediction tasks.**

***Index Terms*—emotion recognition, sentiment prediction, small language models, polish stock market**

I. Introduction

Predicting stock market movements is a complex and widely studied problem, with many researchers and institutions leveraging investor opinions from online sources to gain an advantage. While such data provides valuable insights, the noisy nature of financial markets makes it difficult for models to reliably extract meaningful features from raw text. Emotion recognition has emerged as a promising method to enhance stock prediction accuracy by structuring and organizing textual information in a more useful way [1].

In recent years, pre-trained transformer models have become the state-of-the-art for tasks like emotion recognition. However, their effectiveness is heavily influenced by the relevance of the pre-training data, which introduces a key challenge when fine-tuning for specific applications like stock market sentiment analysis. This problem is further complicated by the scarcity of domain-specific models for languages like Polish, which lack publicly available resources tailored to financial texts.

This paper explores the application of small, pre-trained language models for emotion recognition in Polish stock market-related texts. Specifically, we fine-tune transformers on business articles concerning WIG20, the Polish equivalent of the S&P 500, to classify emotions such as anger, anticipation, joy, sadness, and trust. Given the challenges posed by the scarcity of Polish-specific models, we evaluate several transformer architectures, compare their performance, and address key issues like class imbalance. Our findings demonstrate that the choice of pre-trained model significantly affects performance, with the Polish RoBERTa model achieving the best results.

The key contributions of this paper are:

- Evaluation of small language models for emotion classification in Polish stock market texts.
- Detailed comparison of multiple transformer architectures on a scarce-resource language task.
- Insights into overcoming class imbalance and the potential benefits of domain-specific pre-training [2].

II. Related Work

Several studies have examined the relationship between various factors and stock price movements, with the analysis of online texts emerging as a popular approach [3], [4]. It is widely recognized that emotions significantly influence decision-making, especially in contexts involving risk and uncertainty, such as the stock market [5]. As a result, gauging public emotions has become a logical method for predicting stock market trends. Everyday phenomena, such as weather, have been shown to impact the moods of large groups, with effects on the stock market being observed when moods influenced by weather are misattributed as information [6], [7]. Individuals in positive moods are generally more susceptible to allowing insignificant factors to affect their decisions, which extends to stock market behaviors [8], [9]. Beyond misattribution, emotions directly expressed about companies are also linked to stock price movements [10]. However, analyzing such emotional expressions presents challenges, particularly in selecting appropriate text features. Although models can be trained to extract features from raw text, this approach is less effective in the noisy stock market environment [1].

A. Sentiment and Emotions

One of the simplest features for stock market prediction is measuring the amount of traffic a company generates, which has shown to improve prediction accuracy [11]. A more sophisticated approach is labeling texts with sentiment and aggregating this information into distinct features, as seen in [12]. Sentiment extraction from text is popular due to the availability of pre-built models, and some works focus on categorizing texts into stock-related labels such as "sell," "hold," or "buy" [13]. Another prominent method for stock market prediction involves labeling texts with emotions.

2375-9259/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDMW65004.2024.00065

Different studies select various emotions depending on the type of text and the company in focus [14]. Analyzing social media, news, and blogs has become increasingly common in stock market prediction [15], with emotional analysis being particularly effective in markets with a higher number of individual investors, such as China [16]. The best-performing emotions vary by stock market [17].

A common emotional taxonomy comes from Ekman, who identified six basic emotions: anger, surprise, disgust, enjoyment, fear, and sadness [18], [19]. This approach was applied in [14], which found happiness to be positively correlated with IPO returns and fear to predict price movements. Another widely used emotion set comes from Plutchik, who identified four contrasting pairs: joy vs. sadness, anger vs. fear, trust vs. disgust, and surprise vs. anticipation [20]. Studies like [10] and [17] have used these emotions to show that anger, trust, and fear are particularly relevant in the Polish stock market.

Different works use varied combinations of emotions, such as love [21], anxiety, and calmness [22], or stress and gloom [23]. However, there is no clear consensus on which emotions are most predictive. Studies like [23] show significant effects of fear and stress, while others, such as [22], highlight that combining emotions with sentiment analysis yields better results. For instance, joy, anger, and fear were the most effective emotions in predicting stock prices following earnings announcements [24], and facial expression analysis also links fear and positive valence with market prices [25].

B. Small Language Models

Sentiment features are widely used in stock market prediction, largely due to the availability of ready-to-use models. For example, [12] applied a term frequency-inverse document frequency (TF-IDF) approach to news articles about the Taiwan 50 index, showing that customized lexicons for each stock performed better than a universal lexicon. A similar method was tested on WIG20, with positive results for the Polish market [26]. Sentiment dictionaries, such as the Harvard IV-4 and Loughran McDonald financial sentiment dictionaries, have also been used in stock market analysis [27], as well as emotion dictionaries like Google's Profile of Mood States (GPOMS) [28]. Other popular dictionary-based sentiment classifiers include VADER [29], TextBlob, and Flair [30], which have been applied in studies like [31]. Some works use bag-of-words or more advanced methods like convolutional neural networks (CNNs) with Word2vec embeddings to classify documents [16].

While these approaches demonstrate that incorporating text features can improve price predictions, they do not leverage the latest advances in text classification. The current state-of-the-art (SOTA) method is fine-tuning pre-trained transformer models [32]–[41]. Transformers outperform older feature extraction methods like bag-of-words and Word2vec due to their ability to capture deep semantic features [42]. In stock market settings, transformers have proven superior to recurrent neural networks (RNNs) and CNNs [43]. Although [2] emphasizes that optimal performance is achieved when models are pre-trained on data similar to the task-specific dataset, this is often impractical due to the high cost and data requirements of pre-training. Nonetheless, fine-tuning existing transformer models remains the best strategy for text classification, as demonstrated in sentiment analysis studies [44].

TABLE I
AVERAGED INTER-ANNOTATOR AGREEMENT.

	Documents	Sentences	Comments
PKN ORLEN	0.72	0.69	0.70
PKO BP	0.68	0.77	0.77

C. Summary

In summary, while various methods like RNNs, CNNs, and Word2Vec offer some advantages for stock price prediction, the literature indicates that fine-tuning transformer models is the most effective approach for emotion recognition in stock market contexts. The main challenges in this field are selecting the most appropriate pre-trained model and addressing the issue of class imbalance in the annotated emotions. This work focuses on identifying the best pre-trained transformer model and the optimal strategies for handling class imbalance in the context of Polish stock market articles.

III. STOCKBRIEF DATASET ANNOTATION

This project involved the manual annotation of stock market articles. Annotators were tasked with labeling emotions—joy, trust, surprise, anticipation, sadness, anger, fear, and stress—based on linguistic cues within individual sentences, along with the implied sentiment (neutral, positive, negative, or ambivalent). A similar process was applied to annotate emotions and sentiment at the document level.

A. Work Stages

There were 20 iterations during the project. The work of the annotators was divided into three stages:

1) **Preparation of the guidelines** (iterations 1 to 8): Seven annotators marked the same texts. Based on their work, guidelines for the following stages were refined.
2) **Clarification of terminology and characteristic emotion markers** (iterations 9 to 18): Twelve annotators worked in a 2+1 mode, where two annotators marked the same articles and a super-annotator resolved any inconsistencies between them. Detailed descriptions of individual emotion markers were defined.
3) **Production** (iterations 19 to 20 and corrections of earlier iterations): Annotators worked independently on individual files, while other annotators acted as super-annotators for quality control.

To ensure the quality of the annotation, a Positive Specific Agreement (PSA) [45] threshold of at least 0.65 was set. Two samples were drawn from two domains, one related to PKN ORLEN (energy services) and the other to PKO BP (banking). Each sample contained 50 articles. Table I presents the results.

TABLE II
CONFUSION MATRIX FOR INDIVIDUAL EMOTIONS – A SAMPLE OF 887 ANNOTATED SENTENCES (3RD ITERATION); THE PERCENTAGE OF SENTENCES IN WHICH EACH EMOTION CO-OCCURRED.

	JOY	FEA	ANT	STR	TRU	ANG
SAD	4.99%	6.12%	2.27%	6.80%	1.59%	2.83%
JOY		1.25%	10.88%	1.36%	11.79%	0.79%
FEA			1.25%	3.29%	0.45%	0.45%
ANT				0.34%	9.30%	0.23%
STR					0.57%	1.13%
TRU						0.00%

B. Annotation Guidelines

The guidelines included the following categories:

- **Types of categories** – This section defines the scope of the work, i.e., the annotation of emotions and sentiment for both individual sentences and the entire text.
- **Workflow** – This part outlines the order in which the annotators were to proceed.
- **Labelling perspective** – The perspective taken is that of the local sender of the text, i.e., the author.
- **Discrimination of the main markers of emotion** – Describes the linguistic markers used to identify emotions.
- **Identification of the relationship between emotion and sentiment** – Specific rules were provided to determine the sentiment of individual sentences based on the emotions.
- **Determination of sentiment for the whole text** – Guidelines on how to establish the overall text sentiment.
- **Specification of emotion markers in relation to stock market terminology** – This section classifies stock market-related terms and their corresponding emotions in stock market texts.

The refinement of the guidelines took into account not only the agreement between annotators but also the specific characteristics of particularly difficult documents and the phenomenon of co-occurring emotions. In the latter case, a confusion matrix was used, and annotators were provided with specific guidance on the most commonly overlapping emotions (see example in Table II). This was especially relevant for the dependence of stress on other emotions and the challenges in distinguishing between joy and trust, or joy and anticipation.

IV. STOCKBRIEF DATASET

The Stockbrief dataset consists of Polish articles related to four of the largest companies traded on the Warsaw Stock Exchange: CD Projekt RED (CDR), Powszechna Kasa Oszczędności Bank Polski (PKO), Polski Koncern Naftowy ORLEN (PKN), and KGHM Polska Miedź (KGHM). All these companies are currently part of the WIG20 index. The articles were gathered from business and economic portals such as *Puls Biznesu* and *Parkiet*. The dataset comprises 399 articles and 6861 sentences. Annotations were created from the perspective of the article's author, aiming to capture the emotions that might have been present during the writing process. Sentences were annotated with eight emotions: anger, anticipation, fear, joy, sadness, stress, surprise, and trust. Additionally, three sentiment labels were used: positive, negative, and neutral, where annotating a sentence with both positive and negative sentiment indicates ambivalence.

Documents were annotated with the same labels as sentences, as well as three stock-related variables: sell, keep, and buy. Each text was annotated by up to seven annotators who worked independently, without access to each other's decisions. A label was considered valid if at least two annotators agreed on it. Table III, Figure 3 and Figure 4 show the number of annotations in the final corpus for each label at the document and sentence levels. Figure 1 and Figure 2 illustrate the distribution of annotations corresponding to each company at the document and sentence levels.

TABLE III
NUMBER OF INSTANCES OF DIFFERENT EMOTIONS.

	Number of annotations	
Emotion	Document-level	Sentence-level
Positive	223	2427
Negative	100	1209
Neutral	287	5586
Anger	15	167
Anticipation	123	866
Fear	37	266
Joy	138	1331
Sadness	71	838
Stress	32	332
Surprise	3	98
Trust	108	950
Buy	170	N/A
Keep	387	N/A
Sell	76	N/A
Total	3041	14070

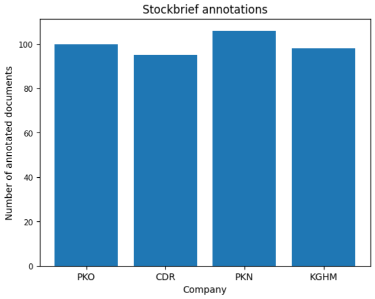

Fig. 1. Number of annotated documents per company.

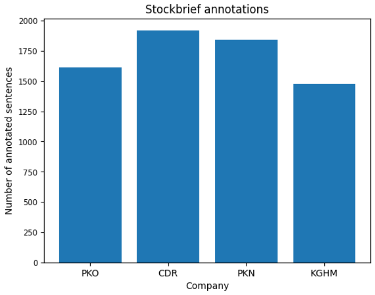

Fig. 2. Number of annotated sentences per company in the Stockbrief dataset.

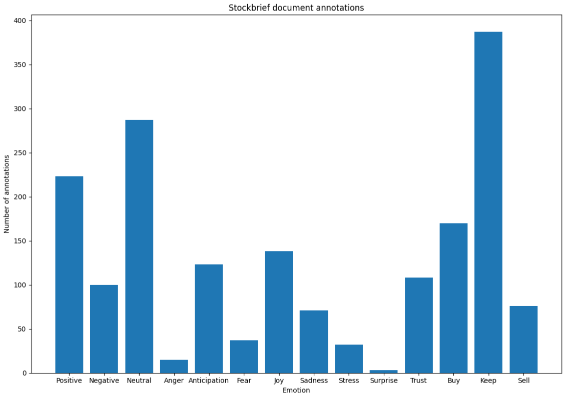

Fig. 3. Number of annotated documents per emotion in the Stockbrief dataset.

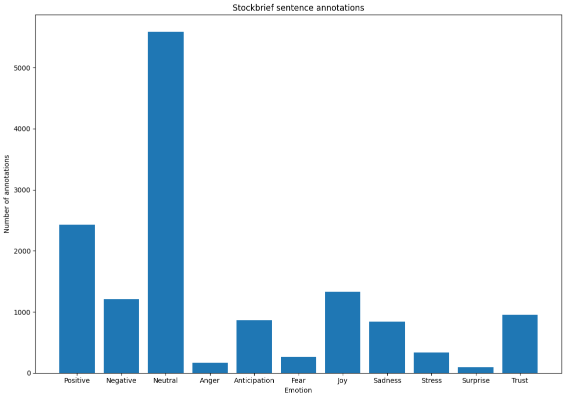

Fig. 4. Number of annotated sentences per emotion in the Stockbrief dataset.

V. Small Language Models

Selecting a pretrained model is a challenging task, as it requires considering multiple factors. One of the most critical decisions is identifying which model's pretraining data most closely aligns with the distribution of the task at hand. Another important consideration is model size, as there is often a trade-off between performance on evaluation metrics and the fine-tuning or inference time. Large models typically require graphics processing units (GPUs) or tensor processing units (TPUs) for fine-tuning, as they tend to run slowly on central processing units (CPUs). Consequently, the hardware must have sufficient video random access memory (VRAM) in the case of GPUs or high-bandwidth memory (HBM) in the case of TPUs, both for fitting the model and for handling the critical batch size and optimizer gradients during training.

It is also essential to account for the distribution and size of the task dataset, as smaller datasets may not provide enough data to fine-tune larger models effectively. In this work, several pretrained models are tested, all of which are based on an encoder-only architecture. This architecture includes a tokenizer and multiple encoder layers. A typical encoder layer consists of a self-attention layer, residual connections, feed-forward layers, and two normalization layers. Additionally, a positional encoding is applied to the text embeddings between the tokenizer and the first encoder layer.

Bidirectional Encoder Representations from Transformer (BERT) [46] is a language representation model based on the encoder part of the transformer architecture. The model is pretrained based on the masked language model (MLM) objective and the next sentence prediction task. The specific version used in this work is the Polish version of the case-sensitive BERT model, with a base size called $pl-bert-cased-base$. This model has a maximum sequence length of 512 tokens and an embedding size of 768. The pretraining of this model was done on a deduplicated version of the Polish subset of Open Subtitles [47], the Polish subset of [48], Polish Parliamentary Corpus [49], and Polish Wikipedia.

DistilBERT [50] is a smaller version of the BERT base multilingual model. It is pretrained using knowledge distillation from the base size BERT multilingual model. The authors claim that this approach allows them to reduce BERT size by 40% and speed it up by 60% while retaining 97% of the previous model's understanding capabilities. The specific version used in this work is the Polish version of the case-sensitive distilBERT model [50], called $distil - bert - pl$. This version gives the same representations as the multilingual distilBERT while having a smaller size. This model has a maximum sequence length of 512 tokens and an embedding size of 768. The BERT base multilingual model [51] has been pre-trained on the top 104 languages with the largest Wikipedia.

Robustly Optimized BERT Pretraining Approach (RoBERTa) [52] differs from BERT by modifying key hyperparameters of the model and removing the next-sentence pretraining objective. The specific version used in this work is the Polish version of the case-sensitive RoBERTa model, called $polish - roberta$ [53]. Both sizes have a maximum sequence length of 514 tokens, with the embedding size being 768 in the base version and 1024 in the large version. The model has been pretrained on a filtered subset of Polish language documents from Common Crawl, Polish Wikipedia, Polish Parliamentary Corpus [49], smaller corpora from CLARIN and OPUS projects, and unspecified Polish books and articles.

XLM-RoBERTa [54] is a multilingual version of A Robustly Optimized BERT Pretraining Approach (RoBERTa) [52]. XLM in the name is derived from a paper describing training methods for cross-lingual language models (XLM) [55]. Similarly to BERT, the model is an encoder part of the transformer model, with the difference being in the training approach. This work tries the base, and the large version of the pre-trained model called $xml - roberta - base$ and $xml-roberta-large$. Both sizes have a maximum sequence length of 514 tokens, with the embedding size being 768 in the base version and 1024 in the large version. It is pre-trained on 2.5TB of filtered CommonCrawl data [56] containing 100 languages.

HerBERT [57] is based on BERT, with the difference being

TABLE IV
PARAMETERS IN DIFFERENT PRETRAINED TRANSFORMERS.

Model	Size [M]	Hidden state	Feed-forward hidden state	Layers	Attention heads	Vocabulary size
bert-base-polish-cased	132	768	3072	12	12	60000
distilbert-base-pl-cased	60	768	3072	12	6	22397
herbert-base-cased	124	768	3072	12	12	50000
herbert-large-cased	355	1024	4096	24	16	50000
polish-roberta-base	124	768	3072	12	12	50001
polish-roberta-large	434	1024	4096	24	16	128001
xlm-roberta-base	278	768	3072	12	12	250002
xlm-roberta-large	559	1024	4096	24	16	250002

TABLE V
DISTRIBUTION OF DOCUMENT LABELS IN DIFFERENT SPLITS.

Emotion	Train	Validation	Test
Negative	72	14	14
Neutral	192	48	47
Positive	159	29	35
Anger	12	2	1
Anticipation	89	17	17
Fear	28	5	4
Joy	102	15	21
Sadness	52	11	8
Stress	21	7	4
Surprise	2	0	1
Trust	76	13	19
Buy	122	25	23
Keep	272	56	59
Sell	55	10	11

that it is pretrained purely on Polish corpora and leverages transferring knowledge from a multilanguage XML-RoBERTa model. The specific version used in this work is the case-sensitive base and large sized models $HerBERT - base$ and $HerBERT - large$. This models have a maximum sequence length of 514 tokens and an embedding size of 768 and 1024 respecti. HerBERT has been pre-trained on six different corpora: CCNet Middle [56], CCNet Head [56], National Corpus of Polish [58], Wikipedia, and Wolne Lektury[1].

A. Model parameters

Pretrained transformers used in this work have different model sizes. Table IV shows the sizes of the models used in this work. Different model sizes trade off the time of fine-tuning and inference with performance the models can achieve.

B. Text and Sentence Classification

A common challenge when using a pretrained transformer is its fixed maximum input sequence length. Frequently, the text to be classified exceeds this limit. The simplest approach employed in this work is to truncate any tokens that exceed the input size, which is applied in document classification. Each input token receives an embedding when the text is passed through the transformer. There are two popular methods for using these embeddings in classification: the first is to utilize only the representation of a classification token, which is always appended at the beginning of the input; the second is to average the embeddings of all tokens. In this work, the first method is used for document classification.

Sentence classification in this work follows the Sequential Sentence Classification approach [59]. In this method, multiple sentences are input into the transformer simultaneously, separated by special separation tokens. The embeddings of these separation tokens are then used as input for the classification head. Multilayer perceptrons (MLP) serve as the classification heads, while BERT is used as the pretrained transformer to generate the embeddings.

As with document classification, a solution is needed when the full document exceeds the maximum sequence length of the transformer. In sentence classification, truncating the entire document is not feasible, as it would remove sentences that require classification. The first step in handling long documents is to split the text if the number of sentences exceeds a predefined limit. This is done by recursively dividing the document in half based on the number of sentences, until each part is smaller than or equal to the defined maximum. Each split is treated as a separate document.

Since the most important context for each sentence typically comes from nearby sentences, this splitting method ensures that each new document contains at least half of the maximum number of sentences. Once divided, each document is tokenized, and each sentence is truncated to a predefined maximum sentence length. If the tokenized document still exceeds the maximum sequence length of the transformer, it is recursively split in half again, based on sentence count, until the size is manageable. After splitting, if a classification token is missing due to the division, it is added at the start. Each split is then treated as a separate document.

VI. EXPERIMENTS

The Stockbrief dataset is split into training, validation, and test sets, based on random selection with proportions of 0.70/0.15/0.15, respectively. For the purpose of sentence classification, the maximum number of sentences that causes splitting a document was set to 10. For the Stockbrief dataset, this creates the label proportions shown in Table V for documents and Table VI for sentences, with the number of instances shown in Table VII.

All training is performed using the AdamW optimization method [60]. In all experiments, the hyperparameters are set as follows: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and the initial value of weight decay, denoted as λ, is set to 0.01. The learning rate, denoted by α, varies between experiments. A warm-up proportion of 10% of the total training steps is used, which is adjusted according to the dataset and the maximum number of epochs. Unless specified otherwise, the maximum number of epochs in the experiments is 100.

Training is conducted with early stopping, meaning that if a metric does not improve after a set number of epochs,

[1] https://wolnelektury.pl/

TABLE VI
DISTRIBUTION OF SENTENCE LABELS IN DIFFERENT SPLITS.

Emotion	Train	Validation	Test
Negative	869	192	148
Neutral	3882	833	871
Positive	1776	301	350
Anger	114	26	27
Anticipation	649	105	112
Fear	198	39	29
Joy	979	154	198
Sadness	606	133	99
Stress	241	53	38
Surprise	64	16	18
Trust	683	129	138

TABLE VII
NUMBER OF EXAMPLES IN DIFFERENT SPLITS.

Split	Documents	Sentences
Train	279	4806
Validtion	60	1024
Test	60	1031

training is halted. The patience parameter for early stopping is set to 25 epochs for all models. The evaluation metric used for early stopping across all experiments is the F_{1micro} score. The output from each classification head is passed through a sigmoid function, after which it is compared to a threshold of 0.5. A dropout layer [61] is applied between the transformer embedding and the classification head. In all experiments, dropout is set to 10% during training.

A. Small Language Models Comparison

To compare the different pretrained models, we use focal loss with additional class weights. The class weights are calculated based on the training set. These weights for the Stockbrief dataset are shown in Table VIII. The learning rate used for both the classification head and fine-tuning the transformer in all experiments in this section is set to 1×10^{-5}. The base versions of the pretrained models use a batch size of 28, while the large models use a batch size of 8.

Table IX compares different hyperparameter values α for document classification using weighted focal loss. The highest F1-macro, F1-micro, and accuracy have been achieved with a value of 0.75, meaning that the loss was greater when considering the positive class for each label.

TABLE VIII
CLASS WEIGHTS. NOT APPLICABLE (N/A) MEANS THIS LABEL DOES NOT EXIST IN THE SENTENCE CLASSIFICATION.

Label	Documents	Sentences
Positive	6.89	4.66
Negative	16.42	10.58
Neutral	5.53	1.59
Anger	103.50	87.25
Anticipation	13.09	14.50
Fear	43.79	49.81
Joy	11.29	9.28
Sadness	23.12	15.60
Stress	58.71	40.75
Surprise	626.00	156.20
Trust	15.50	13.73
Buy	9.28	N/A
Keep	3.61	N/A
Sell	21.80	N/A

TABLE IX
DOCUMENT CLASSIFICATION [%] FOR DIFFERENT α USING WEIGHTED FOCAL LOSS ON *polish-roberta-large*, LEARNING RATE: 1E-5.

α	γ	F1-macro	F1-micro	Accuracy
0.25	2	48.55	70.33	**82.02**
0.50	2	57.58	72.98	81.67
0.60	2	57.41	72.95	81.19
0.75	2	**59.39**	**74.54**	**82.02**
0.90	2	57.79	72.59	79.05
0.95	2	53.53	69.21	74.05

TABLE X
DOCUMENT CLASSIFICATION [%] USING WEIGHTED FOCAL LOSS; $\alpha = 0.75$, $\gamma = 2$, LEARNING RATE: 1E-5. EMOTION LABELS: F1-SCORE.

Measure	bert-base-polish-cased	distilbert-base-pl-cased	herbert-base-cased	herbert-large-cased	polish-roberta-base	polish-roberta-large	xlm-roberta-base	xlm-roberta-large
F1-macro	47.71	43.69	48.05	51.66	57.81	**59.39**	46.55	50.14
F1-micro	69.64	64.45	68.87	72.89	71.88	**74.54**	63.86	71.64
Accuracy	78.10	70.71	76.43	**82.38**	77.74	82.02	67.26	79.64
Macro precision	46.67	35.83	41.08	49.70	47.74	**54.37**	36.78	45.74
Macro recall	54.76	59.59	60.84	57.93	**78.71**	70.97	70.64	58.35
Positive	72.73	65.12	71.79	73.24	81.58	**83.33**	72.34	73.97
Negative	55.17	41.18	50.00	51.61	**70.59**	60.61	54.55	55.56
Neutral	87.85	87.85	87.85	**88.24**	86.79	86.79	87.85	87.85
Anger	00.00	00.00	00.00	**50.00**	**50.00**	33.33	00.00	00.00
Anticipation	61.90	48.15	62.22	59.26	60.87	**70.27**	43.24	66.67
Fear	00.00	**40.00**	37.50	22.22	14.29	33.33	35.29	15.38
Joy	50.85	52.05	**60.00**	57.14	57.63	52.83	50.00	58.18
Sadness	**60.00**	36.36	36.36	55.56	59.26	40.00	52.17	47.62
Stress	40.00	00.00	20.00	00.00	66.67	**85.71**	15.38	33.33
Surprise	**00.00**	**00.00**	**00.00**	**00.00**	**00.00**	**00.00**	**00.00**	**00.00**
Trust	50.00	49.12	47.83	43.75	**56.14**	47.37	47.37	52.63
Sell	35.71	35.90	40.00	55.17	45.45	**72.00**	40.00	50.00
Keep	**99.16**	**99.16**	**99.16**	**99.16**	**99.16**	**99.16**	**99.16**	**99.16**
Buy	54.55	56.79	60.00	**67.86**	60.87	66.67	54.32	61.54

The experiment results for the document classification using focal loss are presented in Table X. The hyperparameters of focal loss used are set to 2 for the focusing parameter and 0.75 for the balance parameter, as suggested in Table IX.

Table XI compares different hyperparameter values α for sentence classification using weighted focal loss. The highest F1-macro has been achieved by a value of 0.75, meaning that the loss was greater when considering the positive class for each label, while the highest F1-micro and accuracy were

TABLE XI
SENTENCE CLASSIFICATION [%] FOR DIFFERENT α USING WEIGHTED FOCAL LOSS ON *polish-roberta-large*, LEARNING RATE: 1E-5.

α	γ	F1-macro	F1-micro	Accuracy
0.25	2.00	50.93	**76.85**	**91.66**
0.50	2.00	51.39	75.84	91.18
0.60	2.00	52.24	75.72	91.06
0.75	2.00	**52.87**	76.60	91.18
0.90	2.00	51.69	75.67	90.44
0.95	2.00	52.31	74.89	89.93

TABLE XII
SENTENCE CLASSIFICATION [%] USING WEIGHTED FOCAL LOSS; $\alpha = 0.75, \gamma = 2$, LEARNING RATE: 1E-5. EMOTION LABELS: F1-SCORE.

Measure	bert-base-polish-cased	distilbert-base-pl-cased	herbert-base-cased	herbert-large-cased	polish-roberta-base	polish-roberta-large	xlm-roberta-base	xlm-roberta-large
F1-macro	47.28	30.70	35.01	30.18	**56.71**	52.87	48.25	49.94
F1-micro	70.89	52.03	56.24	57.41	74.04	**76.60**	68.85	74.93
Accuracy	88.89	73.59	77.15	79.84	89.92	**91.18**	86.77	90.64
Macro precision	46.62	23.16	26.98	25.27	51.67	**53.58**	41.48	50.05
Macro recall	52.37	55.71	60.07	50.28	**67.70**	59.37	65.27	56.45
Positive	66.60	50.73	55.13	50.21	70.84	72.05	66.34	**72.59**
Negative	59.06	36.44	39.67	43.99	**66.41**	64.74	58.97	64.21
Neutral	91.44	90.84	90.98	91.20	**92.55**	92.46	91.78	91.15
Anger	26.50	12.62	28.31	08.27	**41.69**	22.19	32.78	15.43
Anticipation	59.35	30.27	36.36	27.03	62.30	**62.91**	54.32	57.49
Fear	27.94	15.22	13.03	09.43	38.34	**39.29**	24.64	34.32
Joy	50.49	37.74	43.97	22.41	54.91	**59.89**	54.06	59.21
Sadness	47.22	23.16	28.38	30.62	53.62	54.04	45.96	**55.37**
Stress	19.01	13.41	11.56	10.75	**45.79**	31.88	37.49	31.55
Surprise	20.29	01.46	06.58	08.37	**41.40**	22.22	12.68	12.37
Trust	52.19	25.76	31.13	29.67	55.93	**59.89**	51.70	55.67

achieved by a value of 0.25 when greater loss corresponded to negative classes.

The experimental results for sentence classification using focal loss are presented in Table XII. The focal loss hyperparameters are set to a focusing parameter of 2 and a balance parameter of 0.75, as recommended in Table XI. The best overall performance in terms of F1-micro and F1-macro was achieved by the *polish-roberta-large* model, highlighting its effectiveness in capturing sentence-level emotions. On the other hand, the highest accuracy was obtained by the *herbert-large-cased* model, which demonstrates that model choice can significantly impact different evaluation metrics.

Interestingly, none of the models were able to effectively capture the *surprise* category, which is the rarest label in the dataset. This suggests that the limited representation of *surprise* in the data makes it a particularly challenging emotion to classify. Future work might explore strategies such as data augmentation or oversampling to improve the performance on this rare label.

VII. CONCLUSIONS AND FUTURE WORK

The results of this study show that different sources of text used for pre-training have a significant impact on the performance of pretrained transformer models. The findings underscore the importance of carefully selecting a pretrained model to achieve optimal results for a given dataset. This work has explored the selection of models for emotion classification of investor opinions in Polish WIG20 stock-related articles, aimed at improving stock market predictions. Despite both sentence and document classifications being applied to the same texts, different models yielded varying, and in some cases significantly different, performances. The experiments conducted suggest that the *polish-roberta-large* model provides the best pre-training for capturing investor emotions in Polish stock market-related articles, outperforming other models in both document and sentence classification tasks.

Several untested aspects remain for future exploration. One prominent area for improvement is the inclusion of metadata, such as stock names, dates, and text sources, when feeding the data to the transformer model. Additionally, performing a pre-training step on unannotated Polish stock-related articles could likely improve the model's performance.

ACKNOWLEDGMENTS

This work was financed by (1) the National Science Centre, Poland, project no. 2021/41/B/ST6/04471; (2) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; (3) the Polish Ministry of Education and Science within the programme "International Projects Co-Funded"; (4) CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Minister of Science under the programme: "Support for the participation of Polish scientific teams in international research infrastructure projects", agreement number 2024/WK/01; (5) the European Union under the Horizon Europe, grant no. 101086321 (OMINO). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them.

REFERENCES

[1] A. Zaichenko, A. Kazakov, E. Kovtun, and S. Budennyy, "The battle of information representations: Comparing sentiment and semantic features for forecasting market trends," *arXiv preprint arXiv:2303.14221*, 2023.

[2] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[3] R. Mao, K. Du, Y. Ma, L. Zhu, and E. Cambria, "Discovering the cognition behind language: Financial metaphor analysis with metapro," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 1211–1216.

[4] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–42, 2024.

[5] G. F. Loewenstein, E. U. Weber, C. K. Hsee, and N. Welch, "Risk as feelings." *Psychological bulletin*, vol. 127, no. 2, p. 267, 2001.

[6] H.-C. Shu and M.-W. Hung, "Effect of wind on stock market returns: evidence from european markets," *Applied Financial Economics*, vol. 19, no. 11, pp. 893–904, 2009.

[7] W. N. Goetzmann, D. Kim, A. Kumar, and Q. Wang, "Weather-induced mood, institutional investors, and stock returns," *The Review of Financial Studies*, vol. 28, no. 1, pp. 73–111, 2015.

[8] D. M. Mackie and L. T. Worth, "Feeling good, but not thinking straight: The impact of positive mood on persuasion," in *Emotion and social judgments*. Garland Science, 2020, pp. 201–219.

[9] H.-C. Shu, "Investor mood and financial markets," *Journal of Economic Behavior & Organization*, vol. 76, no. 2, pp. 267–282, 2010.

[10] A. Mudinas, D. Zhang, and M. Levene, "Market trend prediction using sentiment analysis: lessons learned and paths forward," *arXiv preprint arXiv:1903.05440*, 2019.

[11] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Systems with Applications*, vol. 79, pp. 153–163, 2017.

[12] M.-Y. Chen, C.-H. Liao, and R.-P. Hsieh, "Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach," *Computers in Human Behavior*, vol. 101, pp. 402–408, 2019.

[13] A. Derakhshan and H. Beigy, "Sentiment analysis on stock social media for stock price movement prediction," *Engineering Applications of Artificial Intelligence*, vol. 85, pp. 569–578, 2019.

[14] D. Vamossy and R. Skog, "Emtract: Investor emotions and market behavior," *arXiv preprint arXiv:2112.03868*, 2021.

[15] O. Bustos and A. Pomares-Quimbaya, "Stock market movement forecast: A systematic review," *Expert Systems with Applications*, vol. 156, p. 113464, 2020.

[16] N. Jing, Z. Wu, and H. Wang, "A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction," *Expert Systems with Applications*, vol. 178, p. 115019, 2021.

[17] D. H. Steyn, T. Greyling, S. Rossouw, J. M. Mwamba *et al.*, "Sentiment, emotions and stock market predictability in developed and emerging markets," Global Labor Organization (GLO), Tech. Rep., 2020.

[18] P. Ekman, "Are there basic emotions?" *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992.

[19] J. Kocoń, "Deep emotions across languages: A novel approach for sentiment propagation in multilingual wordnets," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2023, pp. 744–749.

[20] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[21] S. Saurabh and K. Dey, "Unraveling the relationship between social moods and the stock market: Evidence from the united kingdom," *Journal of Behavioral and Experimental Finance*, 2020.

[22] S. R. Jammalamadaka, J. Qiu, and N. Ning, "Predicting a stock portfolio with the multivariate bayesian structural time series model: Do news or emotions matter?" *International Journal of Artificial Intelligence*, vol. 17, no. 2, pp. 81–104, 2019.

[23] J. Griffith, M. Najand, and J. Shen, "Emotions in the stock market," *Journal of Behavioral Finance*, vol. 21, no. 1, pp. 42–56, 2020.

[24] D. F. Vamossy, "Investor emotions and earnings announcements," *Journal of Behavioral and Experimental Finance*, vol. 30, p. 100474, 2021.

[25] A. Breaban and C. N. Noussair, "Emotional state and market behavior," *Review of Finance*, vol. 22, no. 1, pp. 279–309, 2018.

[26] E. Niedzielska, "Using text analysis for evaluating the behaviour of rates of return from the wig20 index," *Informatyka Ekonomiczna*, 2020.

[27] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 14–23, 2014.

[28] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[29] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *AAAI conference on web and social media*, 2014.

[30] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1638–1649.

[31] P. Prajapati, "Predictive analysis of bitcoin price considering social sentiments," *arXiv preprint arXiv:2001.10343*, 2020.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[33] J. Baran and J. Kocoń, "Linguistic knowledge application to neurosymbolic transformers in sentiment analysis," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 395–402.

[34] W. Korczyński and J. Kocoń, "Compression methods for transformers in multidomain sentiment analysis," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 419–426.

[35] P. Miłkowski, M. Gruza, P. Kazienko, J. Szołomicka, S. Woźniak, and J. Kocoń, "Multiemo: language-agnostic sentiment analysis," in *International Conference on Computational Science*. Springer, 2022, pp. 72–79.

[36] P. Miłkowski, K. Karanowski, P. Wielopolski, J. Kocoń, P. Kazienko, and M. Zięba, "Modeling uncertainty in personalized emotion prediction with normalizing flows," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2023, pp. 757–766.

[37] S. Woźniak and J. Kocoń, "From big to small without losing it all: Text augmentation with chatgpt for efficient sentiment analysis," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2023, pp. 799–808.

[38] J. Kocoń, J. Baran, K. Kanclerz, M. Kajstura, and P. Kazienko, "Differential dataset cartography: Explainable artificial intelligence in comparative personalized sentiment analysis," in *International Conference on Computational Science*. Springer, 2023, pp. 148–162.

[39] J. Kocon, J. Baran, and K. Kanclerz, "Multi-modal personalized hate speech analysis using differential dataset cartography." in *DE-FACTIFY@ AAAI*, 2023.

[40] M. Kochanek, I. Cichecki, O. Kaszyca, D. Szydło, M. Madej, D. Jędrzejewski, P. Kazienko, and J. Kocoń, "Improving training dataset balance with chatgpt prompt engineering," *Electronics*, vol. 13, no. 12, p. 2255, 2024.

[41] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang, N. D. Lane, and M. Xu, "Small language models: Survey, measurements, and insights," *arXiv preprint arXiv:2409.15790*, 2024.

[42] Z. Shuai, D. Xiaolin, Y. Jing, H. Yanni, C. Meng, W. Yuxin, and Z. Wei, "Comparison of different feature extraction methods for applicable automated icd coding," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 11, 2022.

[43] J. Liu, H. Lin, X. Liu, B. Xu, Y. Ren, Y. Diao, and L. Yang, "Transformer-based capsule network for stock movement prediction," in *Proceedings of the first workshop on financial technology and natural language processing*, 2019, pp. 66–73.

[44] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM computing surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[45] G. Hripcsak and A. S. Rothschild, "Agreement, the F-Measure, and Reliability in Information Retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, 05 2005.

[46] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[47] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[48] M. Bañón, P. Chen *et al.*, "Paracrawl: Web-scale acquisition of parallel corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[49] M. Ogrodniczuk, "Polish parliamentary corpus," in *Proceedings of the LREC 2018 workshop ParlaCLARIN: creating and using parliamentary corpora*, 2018, pp. 15–19.

[50] A. Abdaoui, C. Pradel, and G. Sigel, "Load what you need: Smaller versions of mutlilingual bert," in *SustaiNLP / EMNLP*, 2020.

[51] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962v2*, 2019.

[52] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[53] S. Dadas, M. Perełkiewicz, and R. Poświata, "Pre-training polish transformer-based language models at scale," in *Artificial Intelligence and Soft Computing: 19th International Conference*. Springer, 2020.

[54] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[55] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[56] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," *arXiv preprint arXiv:1911.00359*, 2019.

[57] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik, "HerBERT: Efficiently pretrained transformer-based language model for Polish," in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, 2021.

[58] A. Przepiórkowski, "Narodowy korpus języka polskiego," Naukowe PWN, 2012.

[59] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. S. Weld, "Pretrained language models for sequential sentence classification," *arXiv preprint arXiv:1909.04054*, 2019.

[60] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[61] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.