Deciphering Sentiments: A Comparative Study of LSTM and BERT Models in Financial Text Analysis

Adjlan Aymen Abdeldjalil
Attou Yassine
Boucenna Abderrahmane
Ghodbane Youcef Islam
Ecole nationale Supérieure d'Informatique (ESI)
Algiers, Algeria
{ja_adjlane, jy_attou, ja_boucenna, jy_ghodbane}@esi.dz

***Abstract*—In the contemporary business landscape, the significance of sentiment analysis has escalated for enterprises seeking to gather user feedback on their products and services. The integration of this analytical tool into mobile applications necessitates a more compact model, while its conjunction with search engines demands enhanced speed. In response to these challenges, we present a comparative analysis of LSTM and BERT models for financial sentiment analysis. It examines the effectiveness of these machine learning approaches in classifying textual sentiments into negative, neutral, and positive categories. The BERT model, with its deep learning architecture, demonstrated superior performance over LSTM and the baseline Multinomial Naive Bayes classifier. The findings highlight the advanced capabilities of transformer-based models in processing complex language data and suggest directions for future research in model enhancement and application.**

***Index Terms*—sentiment analysis, NLP, stemming, naive Bayes, BERT, LSTM**

I. Introduction

Over the past few years, sentiment analysis has attracted significant interest and has become a crucial instrument across different sectors. Its extensive use can be credited to its effectiveness in collecting and evaluating user sentiments on a wide range of topics. This analytical method has become a preferred choice not just for businesses but also for democratic government bodies aiming to gain valuable insights into public opinions.

There are many possible areas of improvement for a sentiment analysis models including but not limited to: reducing model size to optimize its performance on less powerful devices as well as boosting its precision and recall rendering it more useful for decision making purposes, we can also work on minimizing testing time as that remains an important factor for any market research tool. In resource-constrained settings, such as mobile applications and edge devices, the necessity for compact models is underscored by limitations in storage and computational capabilities. At the same time improving one if the characteristics of our model must not come on the expense of the rest.The challenge lies in achieving synergy across various performance indicators, ensuring that improvements in one dimension do not compromise the overall effectiveness of the model. This holistic approach requires thoughtful consideration of application-specific needs and careful evaluation of trade-offs to create models that are not only optimized for specific criteria but are also robust, versatile, and well-aligned with the overarching objectives of their intended use cases.

II. Related works

In this section, we will present two main approaches to sentiment analysis. The two of them are widely used till nowadays.

A. Rule-based approach

Rule-based approach entails the establishment of a predefined set of rules or patterns for discerning sentiment expressions. Typically, these methods leverage sentiment lexicons or dictionaries that feature words annotated with their respective sentiment polarity. The sentiment of a given text is ascertained by aggregating the sentiment scores associated with individual words. An exemplar in this domain is VADER [1], a method that amalgamates lexical features with consideration for five overarching rules encapsulating grammatical and syntactical conventions for expressing and emphasizing sentiment intensity.

Another notable rule-based sentiment analysis algorithm is presented in [2], specifically tailored for polarity classification of financial news articles. This system employs a pre-existing polarity lexicon to classify financial news articles into positive or negative categories. Sentiment composition rules are applied to ascertain the polarity of each sentence within the news article, while the Positivity/Negativity ratio (P/N ratio) is utilized to compute the overall sentiment values of the news article's content.

In a similar vein, [3] introduces two methodologies, with one adopting a lexicon-based approach. The lexicon undergoes augmentation through the annotation of specific words as positive or negative. Subsequently, Levenshtein distance is invoked to gauge the similarity between two words in terms of characters. The lexicon is further expanded by assimilating words identified as similar through this process. The scoring of each sentence is contingent upon the consideration of the words it encompasses.

B. ML-based approach

The machine learning (ML)-based approach involves training a model on either a labeled or unlabeled dataset, enabling it to discern the sentiment class. In the method proposed by [3], an alternative technique involves feeding word scores into a machine learning algorithm for sentiment classification. A parallel study conducted by [4] takes a distinctive approach, representing word polarity not as discrete values (1 or -1) but as continuous scores. In this work, the authors translated an English lexicon to Arabic, preserving the associated scores.

Additionally, the research of [5] introduces a machine learning-based methodology for summarizing user opinions expressed in reviews. This approach incorporates sentiment knowledge to calculate sentence sentiment scores, utilizing various strategies to address challenges like sentiment shifters, sentence types, and word coverage limits. The method also employs a word embedding model inspired by deep learning to comprehend word meanings and semantic relationships, extracting vector representations for each word. Statistical and linguistic knowledge further contribute to determining salient sentences.

In recent developments, deep learning techniques are gaining prominence in sentiment analysis. [6] propose a hybrid approach, combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) models with Doc2vec embedding. This configuration is tailored for opinion analysis in lengthy texts, reflecting the evolving trend of incorporating deep learning methodologies in sentiment analysis research.

III. Method Overview

Our sentiment analysis approach leverages two advanced machine learning models: LSTM (Long Short-Term Memory) and BERT (Bidirectional Encoder Representations from Transformers). These models are applied to a dataset of financial sentiments, aiming to classify sentences into various sentiment categories. In order to perform sentiment analysis on our dataset we employed a standard text processing pipeline that goes through 3 phases: preprocessing, sentence representation and classification to get the final results. However each of the 2 models that we used deal with these 3 phases differently ; for the LSTM model we implement each of the phases explicitly but the BERT method has its own ready to use pipeline.

A. LSTM with Pretrained Google Word2Vec Model

The LSTM model architecture in our study is integrated with a pretrained Google Word2Vec model for enhanced word representations. Here this model does the following tasks: preprocessing, and classification. Each of these phases is systematically executed in a sequential manner, as illustrated in Figure 1.

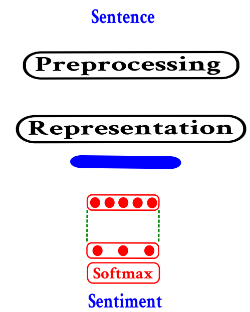

Fig. 1. Summary of tasks in the method

1) Data Preprocessing: The preprocessing involves:

- Tokenization of text data.
- Utilizing Google's Word2Vec for word embeddings.
- Padding sequences for uniform input length.

2) Sentence representation:

3) Model Architecture: The LSTM model comprises:

- An Embedding layer, utilizing Word2Vec weights.
- An LSTM layer with 300 units to process text sequences.
- Dense layers with dropout for regularization.
- A softmax output layer for classification.

4) Classification: The LSTM (Long Short-Term Memory) model classifies sentences by processing input sequences through a series of layers. The process begins with the embedding layer, which converts words into numerical vectors using pre-trained Word2Vec weights. The LSTM layer captures sequential dependencies and long-term context in the input sequence, providing the model with the ability to understand the temporal relationships between words. The output from the LSTM layer is then fed into dense layers with dropout for regularization, enabling the model to learn complex patterns in the data. The final softmax output layer converts the model's output into probabilities for different sentiment classes. During training, the model adjusts its weights to minimize the difference between predicted and actual labels using an optimization algorithm and backpropagation. In inference, the trained model processes new sentences, producing a probability distribution over sentiment classes and selecting the class with the highest predicted probability as the final sentiment classification.

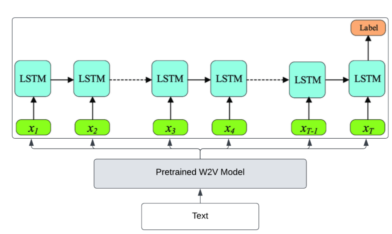

Fig. 2. Lstm architecture

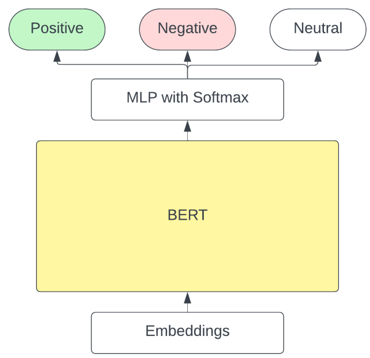

Fig. 3. Bert architecture

B. BERT Model

BERT, known for its deep bidirectional nature, is employed for its superior context understanding. Bert has ready made implementations of the preprocessing, sentence representation and classification

1) Model Architecture: The BERT model structure includes:

- A pretrained BERT layer ('bert-base-uncased').
- A softmax output layer for class probabilities.

Both models, LSTM with Word2Vec and BERT, are designed to capture the nuances and contexts within financial sentiment texts. They represent the state-of-the-art in natural language processing and offer unique approaches to sentiment classification.

IV. Experiment

A. Metrics

To assess the efficacy of our method, we employ recall, precision, and F1-score as articulated in Equations 1 and 2.

$$R = \frac{TP}{TP + FN} \qquad P = \frac{TP}{TP + FP} \tag{1}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{2}$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. Additionally, we incorporate prediction time as a metric for comparing different systems.

B. Data Preparation

We utilize the "Financial sentiment analysis" dataset, comprising 5842 samples categorized into 1852 positive, 3130 neutral, and 860 negative instances. This dataset is partitioned into training (70%) and test (30%) sets.

C. Baseline

In our comparative analysis, the baseline system is defined by a Multinomial Naive Bayes (MNB) model trained with Term Frequency (TF) encoding of sentences. The MNB serves as a standard for comparison due to its simplicity and effectiveness in text classification tasks. This probabilistic model, based on Bayes' theorem, is particularly suitable for high-dimensional datasets and has proven to be a robust classifier for text analysis.

D. Results and Discussion

1) LSTM Model: The LSTM model's performance was benchmarked against the baseline Multinomial Naive Bayes classifier. While the LSTM demonstrated an improved ability to capture nuances in sentence representations, especially in the neutral category, it still showed room for improvement in the negative sentiment category when compared to the baseline MNB model.

2) BERT Model: The BERT model, in comparison to both the LSTM and the baseline MNB, displayed a significant performance leap. Notably, the BERT model's deep understanding of context and language nuances resulted in a substantial increase in precision and recall across all sentiment categories, particularly outshining the baseline in positive sentiment classification.

3) Comparative Analysis: When directly compared, the BERT model outperforms both the LSTM model and the baseline MNB in terms of precision, recall, and F1-score. This superiority can be attributed to BERT's bidirectional architecture and its pre-trained language representations, which are absent in LSTM and the baseline MNB classifier.

Table I presents the test performance of the LSTM and BERT models on the sentiment analysis task. The BERT model demonstrates superior performance across all categories when compared to LSTM, underscoring its effectiveness in understanding the context and nuances in financial sentiment analysis.

Model	Sentiment	Precision	Recall	F1-Score
LSTM	negative	0.52	0.54	0.53
LSTM	neutral	0.73	0.83	0.78
LSTM	positive	0.78	0.58	0.67
BERT	negative	0.52	0.73	0.61
BERT	neutral	0.89	0.75	0.81
BERT	positive	0.82	0.88	0.85

TABLE I
Sentiment classification performance of LSTM and BERT models.

Table II provides a summary of the overall accuracy and macro averages for both models, including the time taken to predict sentiments on the test set.

4) Implications of Results: These results highlight the strengths of transformer-based models like BERT in handling complex natural language processing tasks, outperforming more traditional methods such as LSTM combined with Word2Vec embeddings and the baseline Multinomial Naive

Model	Accuracy	Time (s)
LSTM	0.71	2.2538
BERT	0.78	13.9878

TABLE II
OVERALL PERFORMANCE COMPARISON BETWEEN LSTM AND BERT MODELS.

Bayes classifier. Future research directions may include exploring advanced embeddings for LSTM or combining LSTM and BERT to leverage the strengths of both models.

V. CONCLUSION

In conclusion, this study has demonstrated the efficacy of advanced machine learning models, LSTM and BERT, for sentiment analysis in financial texts. The BERT model, in particular, showcased its superior capability in capturing the nuances and context of language, outperforming both the LSTM model and naive bayes across all sentiment categories. These findings underscore the potential of transformer-based models in natural language processing tasks, offering significant improvements over traditional models such as LSTM with Word2Vec embeddings and the baseline Multinomial Naive Bayes classifier. Future directions may include integrating more sophisticated embeddings or hybrid models to further enhance performance and applicability in real-world sentiment analysis applications.

REFERENCES

[1] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[2] L. Im Tan, W. San Phang, K. O. Chin, and A. Patricia, "Rule-based sentiment analysis for financial news," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2015, pp. 1601–1606.

[3] M. Bettiche, M. Z. Mouffok, and C. Zakaria, "Opinion mining in social networks for algerian dialect," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, J. Medina, M. Ojeda-Aciego, J. L. Verdegay, I. Perfilieva, B. Bouchon-Meunier, and R. R. Yager, Eds. Cham: Springer International Publishing, 2018, pp. 629–641.

[4] I. Guellil, A. Adeel, F. Azouaou, F. Benali, A.-e. Hachani, and A. Hussain, "Arabizi sentiment analysis based on transliteration and automatic corpus annotation," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 335–341. [Online]. Available: https://www.aclweb.org/anthology/W18-6249

[5] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment," *Expert Systems with Applications*, vol. 109, pp. 66–85, 2018.

[6] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A cnn-bilstm model for document-level sentiment analysis," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 832–847, 2019.