CoMFinSe-MusCaAt: Code-Mixed Financial Sentiment Classification via Multi-scale Context-Aware Attention on Low-Resource Language Settings

Ashraf Kamal, Padmapriya Mohankumar, and Vishal Kumar Singh

Abstract Financial sentiment is a key qualitative measure to analyse the opinions and emotions of investors and market participants regarding financial markets and assets. Classifying financial sentiments with linguistic variations provide valuable insights to multi-lingual communities, wherein multiple languages are mainly taken for financial discussions. In this line, this study presents a new approach for financial sentiment classification. We propose a new multi-scale context-aware `CoMFinSe-MusCaAt` model to classify financial sentiment over English, low-resource (Hindi), and code-mixed (Hindi and English) related three datasets. The performance of the proposed `CoMFinSe-MusCaAt` model shows impressive results across all datasets and languages. It also outperforms relevant studies and baseline methods in terms of *F-score* and *Accuracy*.

Keywords Sentiment analysis · Financial sentiment classification · Attention layer · Low-resource language · Code-mixed

1 Introduction

In last two decades, the emergence of the Web has caused large volumes of unstructured data like text, audio, and video [1, 2]. The rapid rise of financial information has also seen in such data [3, 4]. This proliferation of financial information has immense potential to deal with various online social media problems in financial domain, such as financial misinformation [5], financial stance [6], mental health [7], and fake news [8]. Researchers consider financial textual information in various applications,

A. Kamal · P. Mohankumar · V. K. Singh (✉)
PayPal, Chennai, India
e-mail: vishalksingh@paypal.com

A. Kamal
e-mail: askamal@paypal.com

P. Mohankumar
e-mail: pamohankumar@paypal.com

A. Swaroop et al. (eds.), *Proceedings of Data Analytics and Management*, Lecture Notes in Networks and Systems 1297, https://doi.org/10.1007/978-981-96-3352-4_27

such as information retrieval, text summarisation, recommendation systems, and sentiment classification. Recently, sentiment classification has become a vital tool to receive insights from numerous financial information in terms of articles, news, stock price etc.

Financial sentiment has a significant impact on market movements and it leads to market bubbles or crashes if sentiment becomes excessively optimistic or pessimistic. Investors and analysts closely monitor sentiment indicators to gauge market sentiment and inform their investment decisions [10]. Financial sentiment classification (FSC) is an important measure for analysing investor opinions. Moreover, analysing and classifying of sentiments in code-mixed and low-resource languages help investors to identify emerging trends, gain insights into the behaviour of different market segments and demographics, and allow investors from various linguistic backgrounds to access relevant market insights and participate in financial markets more effectively. Also, code-mixed and low-resource languages have cultural nuances and sentiments that might not be captured in mainstream financial analysis. Computational linguistics mainly consider FSC in English language, but less research exploration has been found in code-mixed and low-resource languages. Hence, considering above facts, classification of financial sentiment in code-mixed and low-resource languages is the need of the hour and worth research exploration task.

1.1 Our Contributions

The classification of sentiment plays a vital role by providing important insights in various domains using different linguistic settings, including finance. Considering this, we demonstrate a problem of FSC in English, code-mixed, and low-resource (Hindi) languages. To the best of our knowledge, this is a new study using this research direction. It is represented as binary classification problem, wherein an input text from above-mentioned languages is classified as either *positive financial sentence* (PFS) or *negative financial sentence* (NFS). In this study, a new model, *CoMFinSe-MusCaAt* is presented. In consists of an input layer which takes input from above-mentioned languages and it is passed to their corresponding end-to-end sequential layers. Each layer has a relevant Transformer-based BERT embeddings followed by local channel (BiLSTM, intra-attention), fused-attention, inter-attention layers and dense layers. BiLSTM is used to retrieve latent semantic features in both directions and intra-attention layer is responsible to receive important relevant sentiment information from related tokens with respect to the given language in the input text. Also, the outcome of the intra-attention layers from English and Hindi language sources are merged with fused-attention and passed to inter-attention layer which refers to mainly cross-attention. The main purpose behind is to allow the model to receive the relevant sentiment-based information of code-mixed (English–Hindi) related sequence with the information receive from single language (English and Hindi)-based sequence layers. Further, the outcome of the all intra/inter attention

layers are passed to the corresponding dense layers where activation function is employed to accomplish classifying of input text as either PFS or NFS with respect to the different input languages. The key contributions of this study are given below:

- Introduced a novel FSC problem on code-mixed with low-resource languages.
- Proposed a new model called CoMFinSe-MusCaAt to classify financial sentiment.
- Performed experiments on English, code-mixed (Hindi–English), and Hindi languages for FSC.
- Compared performance evaluation results with relevant studies and baseline methods.

2 Related Work

This section presents the recent works related to FSC problem in textual data. In [10], author considered SemEval-2017 dataset for financial sentiment analysis via knowledge-base and machine learning techniques. In [9], authors highlighted market-derived related financial sentiments. In [11], authors proposed a model called LECN which leverage lexicon supported targeted financial sentiment analysis. In [12], authors considered aspect-based financial sentiment analysis. In [13], authors proposed financial sentiment analysis for stock movement prediction via BERT and CNN, RNN, and LSTM. In [14], authors considered financial news for sentiment analysis. They combined ambiguity and polysemy on augmented imbalanced dataset. They considered CNN along with attention to receive sentiment. In [15], authors considered 10,000 tweets for financial market sentiment. They have taken logistic regression model which outperforms BERT model due to domain specificity. In [16], author proposed FinBERT, a kind of BERT to perform sentiment analysis on several corpora of financial news headlines. In [19], authors considered deep neural network for sentiment analysis. In [20], authors considered sentiment analysis on financial data. In [21], authors considered financial sentiment analysis.

It can be seen that researchers emphasise less focus on FSC problem in the existing studies despite it is an important problem, especially in code-mixed and low-resource language settings. Hence, investigation of FSC is a worth research task to explore.

3 Proposed Approach

This section presents the proposed approach. It constitutes dataset collection and preprocessing followed by the brief discussion on the proposed CoMFinSe-MusCaAt model.

Table 1 Final Statistics of all datasets post pre-processing

Datasets ↓	Positive	Negative	Source lang.	Translated lang.
(DS-1): Malo et al. [17]	570	303	English	Hindi
(DS-2): Mazia et al. [18]	3685	2106	English	Hindi
(DS-3): Code-mixed	3555	1592	English–Hindi	–

3.1 Dataset Collection and Pre-processing

In this study, we consider two publicly available benchmark datasets based on financial sentiment in English and consider only positive and negative instances. Further, we translate these datasets into Hindi through `Deep Translator`[1] API which import `Google Translator`. We leverage ChatGPT 3.5 to generate code-mixed (English–Hindi) dataset.[2]

Further, we perform several pre-processing steps with an aim to clean the collected data. Hence, we have excluded comma, punctuation, more than one dots, question marks, alpha-numeric characters, URLs, mentions, hashtags, exclamation marks, and in the end convert collected data into lower-case form. Table 1 presents the statistics of the dataset post pre-processing.

3.2 Proposed `CoMFinSe-MusCaAt` *Model*

This section presents the newly proposed `CoMFinSe-MusCaAt` model for `FSC` task. Figure 1 presents the architectural work-flow of the proposed model. The complete description of all layers of the proposed model is demonstrated in the following sub-section.

3.2.1 Input Layer

The input layer receives textual data in English, Hindi, and code-mixed (English–Hindi) separately. Words available in the input text are tokenized and allocated with a numerical index value for three input languages. It forms a dictionary accordingly and converted as an input vector, v_i. Thereafter, a fixed-padding of size 64 is applied on each input vector, v_i to maintain the same padding length. Consequently, padded vectors (p_e, p_h, and p_c) are formed of same length and these are passed to their corresponding embeddings in the next layer.

[1] https://deep-translator.readthedocs.io/en/latest/installation.html.

[2] https://chat.openai.com/.

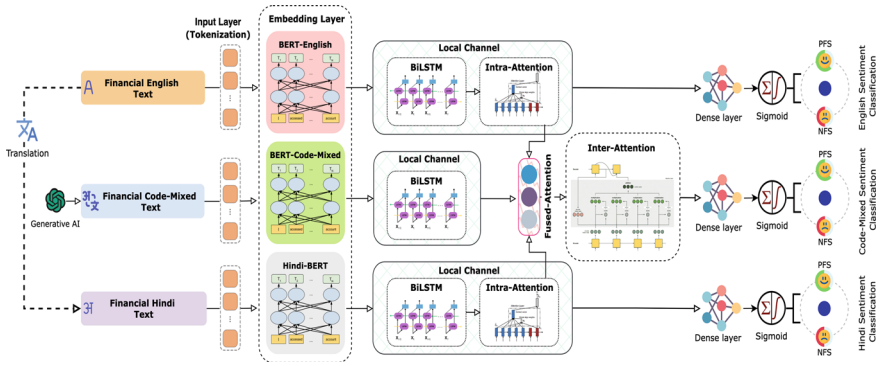

Fig. 1 Architectural work-flow of our proposed `CoMFinSe-MusCaAt` model

3.2.2 Embedding Layer

The embedding layer consists of three embeddings. `BERT` is one of the most recognised pre-trained language models used to receive the input padded vector, p_e which is generated from input text in English. `HindiBERT`,[3] a Hindi BERT model is used for input padded vector, p_h formed from input text in Hindi. Likewise, a BERT code-mixed base model for Hinglish[4] is used for input padded vector, p_c formed from input text in code-mixed (English–Hindi). All embeddings produce relevant and contextual word vector representation based on the financial sentiment through 768-dimensional using `Transformer`. Lastly, the encoded representation of three embeddings are passed to their corresponding next layers.

3.2.3 Local Channels

The local channels in our proposed model consist of two layers (BiLSTM and inter-attention layers) for English and Hindi languages-based input sequences whereas in case of code-mixed only BiLSTM is used.

- **BiLSTM layer**: BiLSTM is one of the popular recurrent neural network which functions in both directions via two gates. In this study, we use one BiLSTM for each languages (English, Hindi, and code-mixed (English–Hindi). It receives the outcome as embedded vector from their corresponding embedding layers of these

[3] https://huggingface.co/l3cube-pune/hindi-bert-v2.

[4] https://huggingface.co/rohanrajpal/bert-base-en-hi-codemix-cased.

languages. It is considered in the proposed model with an aim to produce contextual and latent semantic feature representation of financial sentiment as forward sequences in forward direction and backward sequences in backward direction with 64 neurons. Equation 1 presents a combined representation of BiLSTM operating in both directions.

$$\text{lstm}_i = [\overrightarrow{\text{lstm}_f}, \overleftarrow{\text{lstm}_b}] \tag{1}$$

- **Intra-attention layer**: In this study, intra-attention layer is used in the proposed model to assign different weights to different tokens based on their financial sentiment related information to each other within the English and Hindi languages-based on input sequences. It receives the outcome of the respective BiLSTM layers. It helps the model to capture contextual information efficiently and generate two context vectors, c_e, and c_h for English and Hindi input sequences, respectively and passed to the fused-attention layer.

3.2.4 Fused-Attention Layer

This layer combines the outcome of the two financial sentiment related context-vectors of intra-attention layers on English and Hindi along with output produced from the BiLSTM layer. It leverages relevant contextual information for both languages as two context vectors, c_e, and c_h and generate a fused context vector, c_f.

3.2.5 Inter-attention Layer

Inter-attention layer receives the fused context vector, c_f from the fused-attention layer. It captures important contextual information from two different sequences (i.e., English and Hindi). It allows the proposed model to generate coherent and contextually-mixed relevant outputs by considering the interactions between the different sequences. As a result, it produces aggregated cross-context vectors as c_m.

3.2.6 Dense Layers and Final Classification

Three dense layers followed by `Sigmoid` activation functions are taken into consideration for final `FSC` on English, code-mixed (English–Hindi), and Hindi languages. Dense layers receive the outcome of the two language specific intra-attention layers and one code-mixed related inter-attention layer.

4 Experimental Setup and Results

In this study, all datasets are taken as 80% for training, 20% for testing, and 10% of validation data is taken from the training data to perform experimental tasks. We use `Intel` processing machine, `Ubuntu` OS, 64-GB RAM, and `NVIDIA` GPU. Our model is developed via `PyTorch`,[5] a machine learning framework in `Python`. We use $1e-5$ learning rate, binary cross-entropy, 100 batch-size, 30 epoch, and `adam` optimizer.

4.1 Results and Comparative Analysis

In this section, we discuss the received results of our proposed model and compared with several comparable methods like relevant studies and baseline methods. Table 2 presents the performance evaluation results of our newly proposed `CoMFinSe-MusCaAt` model on two benchmark and newly created datasets for English, low-resource related Hindi, and code-mixed languages. Observe that, our `CoMFinSe-MusCaAt` proposed model shows impressive results in terms of *F-score* and *Accuracy* in all categories of languages across all datasets. Further, it is also observe that, the proposed model outperforms both relevant studies and several baseline models and their combinations. The proposed model receive the highest 0.95 and 0.96 *F-score* and *Accuracy* for code-mixed data. Also, it gives impressive results as 0.90 *F-score* and 0.93 *Accuracy* for low-resource language like Hindi.

These results give a remarkable indication about our proposed model is quite efficient to perform significantly on low-resource and code-mixed languages along with English. It shows that consideration of the context-aware relevant embeddings and fusion of inter and intra-attention layers contribute an important role in enhancing the model performance.

5 Conclusion and Future Works

This study has introduced `FSC` in three languages—English, low-resource (Hindi), and code-mixed (English–Hindi). A novel model called as `CoMFinSe-MusCaAt` is introduced to classify financial sentiment on two benchmark and a new code-mixed datasets across these three languages. Our proposed model has shown remarkable results over all datasets. It has also performed significantly better than compared methods. The `FSC` on figurative language categories like sarcasm where implicit sentiment are involved could be a worth research problem in future. Moreover, exploring `FSC` in multi-modal settings could also be an interesting research direction.

[5] https://pytorch.org/.

Table 2 Performance evaluation results on two benchmark (Malo et al. [17] (DS-1) and Mazia et al. [18] (DS-2)) and code-mixed (DS-3) datasets for English, Hindi, and code-mixed (English–Hindi) languages in terms of *F-score* (F-Sc.) and *Accuracy* (Acc.)

Languages →	English				Hindi				Code-mixed	
Datasets →	DS-1		DS-2		DS-1		DS-2		DS-3	
Methods ↓	F-Sc.	Acc.	F-Sc.	Acc.	F-Sc.	Acc.	F-Sc.	Acc.	F-Sc.	Acc.
Our Proposed model	**0.94**	**0.95**	**0.91**	**0.94**	**0.90**	**0.93**	**0.88**	**0.89**	**0.95**	**0.96**
Priyadarshini and Cotton [19]	0.88	0.86	0.87	0.91	0.88	0.90	0.84	0.85	0.91	0.92
Issam et al. [20]	0.65	0.65	0.64	0.64	0.51	0.65	0.50	0.64	0.69	0.69
BiLSTM	0.79	0.79	0.72	0.72	0.84	0.84	0.70	0.70	0.93	0.93
BiGRU	0.83	0.83	0.74	0.74	0.85	0.85	0.72	0.72	0.92	0.92
CNN + BiLSTM	0.82	0.82	0.73	0.73	0.86	0.86	0.68	0.68	0.93	0.93
CNN + BiGRU	0.81	0.81	0.74	0.74	0.84	0.84	0.69	0.69	0.94	0.94
BERT + BiLSTM	0.91	0.91	0.74	0.74	0.82	0.82	0.73	0.74	0.94	0.94
RoBERTa + BiLSTM	0.84	0.85	0.76	0.77	0.51	0.65	0.59	0.65	0.93	0.93

References

1. Kamal A (2021) A unified data mining approach for detecting figurative language in Twitter (2021). https://shodhganga.inflibnet.ac.in/handle/10603/441185
2. Haswani V, Mohankumar P (2022) Methods to optimize Wav2Vec with language model for automatic speech recognition in resource constrained environment. In: Proceedings of the 19th international conference on natural language processing (ICON). ACL, IIIT Delhi, India, pp 149–153
3. Kamal A, Anwar T, Sejwal VK, Fazil M (2024) BiCapsHate: attention to the linguistic context of hate via bidirectional capsules and hatebase. IEEE Trans Comput Soc Syst (TCSS) 11(2):1781–1792
4. Kamal A, Abulaish M, Jahiruddin (2024) Contextualized satire detection in short texts using deep learning techniques. J Web Eng 23(1):27–52
5. Kamal A, Mohankumar P, Singh VK (2023) Financial misinformation detection via RoBERTa and multi-channel networks. In: Proceeding of 10th international conference on pattern recognition and machine intelligence (PReMI). LNCS. Springer Nature, Switzerland; ISI, Kolkata, India, pp 646–653
6. Singh VK, Mohankumar P, Kamal A (2023) CoMFinSe-MusCaAt: a novel deep learning-based multi-task model for detecting financial stance and sentiment. In: Proceedings of the 14th international conference on computing communication and networking technologies (ICCCNT). IEEE, IIT Delhi, India, pp 1–6
7. Kamal A, Mohankumar P, Singh VK (2022) IMFinE: an integrated BERT-CNN-BiGRU model for mental health detection in financial context on textual data. In: Proceedings of the 19th international conference on natural language processing (ICON). ACL, IIIT Delhi, India, pp 139–148
8. Mohankumar P, Kamal A, Singh VK, Satish A (2023) Financial fake news detection via context-aware embedding and sequential representation using cross-joint networks. In: Proceedings of the 15th international conference on COMmunication systems & NETworkS (COMSNETS). IEEE, Bengaluru, India, pp 780–784
9. Luo L, Ao X, Pan F, Wang J, Zhao T, Yu N, He Q (2018) Beyond polarity: interpretable financial sentiment analysis with hierarchical query-driven attention. In: Proceedings of the IJCAI, pp 4244–4250, Stockholm, Sweden
10. Agarwal B (2023) Financial sentiment analysis model utilizing knowledge-base and domain-specific representation. Multimed Tools Appl 82(6):8899–8920
11. Shang L, Xi H, Hua J, Tang H, Zhou J (2023) A lexicon enhanced collaborative network for targeted financial sentiment analysis. Inf Process Manage 60(2):103187
12. Jangid H, Singhal S, Shah RR, Zimmermann R (2018) Aspect-based financial sentiment analysis using deep learning. In: Proceedings of the the web conference, pp 1961–1966
13. Othan D, Kilimci ZH, Uysal M (2019) Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (BERT) and deep learning models. In: Proceedings of the ICIIT, pp 30–35, Istanbul, Turkey
14. Adhikari S, Thapa S, Naseem U, Lu HY, Bharathy G, Prasad M (2023) Explainable hybrid word representations for sentiment analysis of financial news. Neural Networks 164:115–123
15. Wilksch M, Abramova O (2023) PyFin-sentiment: towards a machine-learning-based model for deriving sentiment from financial tweets. Int J Inf Manage Data Insights 3(1):100171
16. Araci D (2019) Finbert: financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 (2019)
17. Malo P, Sinha A, Korhonen P, Wallenius J, Takala P (2014) Good debt or bad debt: detecting semantic orientations in economic texts. J Assoc Inf Sci Technol 65(4):782–796
18. Maia M, Handschuh S, Freitas A, Davis B, McDermott R, Zarrouk M, Balahur A (2018) WWW' 18 open challenge: financial opinion mining and question answering. In: Proceedings of the web conference, pp 1941–1942
19. Priyadarshini I, Cotton C (2021) A novel LSTM-CNN-grid search-based deep neural network for sentiment analysis. J Supercomput 77(12):13911–13932

20. Issam A, Mounir AK, Saida ElM, Fatna EM (2022) Financial sentiment analysis of tweets based on deep learning approach. Indonesian J Electr Eng Comput Sci 25(3):1759–1770
21. Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM (2018) Big data: deep learning for financial sentiment analysis. J Big Data 5(1):1–25