I.J. Information Engineering and Electronic Business, **2025, 2, 51-67**
Published Online on April 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2025.02.02

Classification of Multilingual Financial Tweets Using an Ensemble Approach Driven by Transformers

Rupam Bhattacharyya*
Department of Information Technology, Gauhati University, Guwahati, PIN-781014, India
E-mail: rupam@gauhati.ac.in
ORCID iD: https://orcid.org/0000-0002-5413-4915
*Corresponding Author

Received: 12 August, 2024; Revised: 10 September, 2024; Accepted: 08 January, 2025; Published: 08 April, 2025

Abstract: There is a growing interest in multilingual tweet analysis through advanced deep learning techniques. Identifying the sentiments of Twitter (currently known as X) users during the IPO (Initial Public Offering) is an important application area in the financial domain. The number of research works in this domain is less. In this paper, we introduced a multilingual dataset entitled as LIC IPO dataset. This work also offers a modified majority voting-based ensemble technique in addition to our proposed dataset. This test-time ensembling technique is driven by fine-tuning of state-of-the-art transformer-based pretrained language models used in multilingual natural language processing (NLP) research. Our technique has been employed to perform sentiment analysis over LIC IPO dataset. Performance evaluation of our technique along with five transformer-based multilingual NLP models over this dataset has been reported in this paper. These five models are namely a) Bernice, b) TwHIN-BERT, c) MuRIL, d) mBERT, and e) XLM-RoBERTa. It is found that our test-time ensemble technique solves this multi-class sentiment classification problem defined over the proposed dataset in a better way as compared to individual transformer models. Encouraging experimental outcomes confirms the efficacy of the proposed approach

Index Terms: IPO, BERT, Multilingual Tweet Processing, Sentiment Analysis, Financial Social Media, Natural Language Processing (NLP)

1. Introduction

Stock markets across the world attract users from different sections of society. Several investment instruments present in those markets have been exercised by them to gain high rewards (monetary profit). In this study, we focus on IPO. In the IPO event, the shares of a company are made available to public investors on the stock market [1]. For a company, launching of IPO is a significant milestone. The company can now have a greater access to capital collected from the public.

Researchers in corporate finance are becoming more interested in the role of social media. Social media is a good place to look for dominant beliefs of retail investors and common people during the launch of a new IPO [2]. Twitter is the most widely used online social media platform. Twitter users share their views / opinions / feelings regarding various financial events in the stock market. Researchers from corporate finance, behavioural economics, and deep learning can contribute to the development of novel applications regarding sentiment analysis of financial tweets. After the launch of a new IPO, such novel research prototypes can help investors to take an immediate intelligent decision after the listing of the company into the stock market. Such decisions may lead to high monetary returns by minimizing the risk.

There are various stock markets across the globe. We focus on Indian stock market. We want to investigate the Life Insurance Corporation of India (LIC) IPO of Indian stock market. LIC is India's largest insurance provider. Users of Indian stock market are huge and several reliable financial applications are further required to empower them with stable regular income. Finding multilingual financial sentiments in tweets during an IPO has several practical applications. Strong market interest and the possibility of larger initial returns for the public could be indicated by positive financial sentiment [3]. Determining the sentiment of these tweets is therefore crucial before the company is listed on a stock exchange. Beyond this, sentiment analysis during an initial public offering (IPO) can aid in forecasting a stock's price movement on the first day of trading. Research has indicated a connection between increased IPO returns

and positive sentiment in tweets. Algorithmic trading strategies can incorporate sentiment analysis data. Based on trends in public sentiment, real-time financial sentiment analysis can offer recommendations for purchasing stocks during an initial public offering (IPO) [4]. To learn more about investor preferences and market positioning, companies can track the conversation surrounding the initial public offerings (IPOs) of their rivals. During the initial public offering (IPO) process, companies can monitor public perception and manage their reputation by using financial sentiment analysis from tweets. Regulators can monitor anomalous sentiment spikes or coordinated negative campaigns using financial sentiment analysis during an initial public offering (IPO) to identify possible market manipulation or fraudulent activities. Such oversight is required because rivals in the industry have the potential to damage a respectable and well-known business during its initial public offering (IPO).

The LIC IPO provides a good experimental platform for analyzing the sentiment of investors during this IPO. The aim of this study is to analyze the tweets of users regarding this IPO before the listing date. There is no existing dataset consisting of such tweets. Therefore, we need to collect those tweets and annotate them appropriately. Analysis of those multilingual financial tweets through state-of-the-art deep learning techniques (present in the NLP literature) is performed in this work. Hindi is one of the official languages of India, with approximately 691 million native speakers. Despite this fact, India is a multilingual country. Indians also use a combination of two languages for their social conversations on Twitter. This process is popularly known code mixing [5]. Since LIC IPO is a significant nation-wide financial event, tweets involving other regional languages can not be ignored. Those multilingual financial tweets along with code-mixed tweets provide a challenging experimental benchmark for the state-of-the-art NLP techniques.

Scientific evolution of transformer-based deep learning models revolutionized the entire NLP area [6]. Novel applications are built by deep learning practitioners due to this evolution. The performance of most of the multilingual models has been shown over specific downstream tasks, which does not include sentiment classification of financial multi-lingual tweets involving IPO. Therefore, our case study is new. This study allows us to explore the possible answers to the following research question (RQ).

RQ: Is it possible for pretrained multilingual language models to give good accuracy over a new downstream task through fine-tuning on less training data?

To figure out the necessity to explore the above research question, readers should be aware of one important dimension of human learning process. Prior experiences are used by human beings while adapting to new and novel scenarios [7, 8]. Pretrained language models are found to be good for different kinds of downstream tasks of NLP. In essence, fine-tuning enables the language model to modify its behaviour and body of knowledge to fit the particular downstream task, even n cases where the task was not explicitly trained on during the pre-training stage. According to the NLP literature, large language models (LLMs) have proven their ability to perform in-context learning (ICL) in a variety of natural language processing tasks [9]. It is claimed that such models perform well over different downstream tasks involving less amount of labelled data [10]. This claim needs further attention from multilingual NLP community. Because, such abilities are present in biological human learning process, and we have to be very careful in validating such claims for language models. Towards this end, we put forward the above research question since our labelled data is less, diverse and multilingual. Furthermore, we all want a data efficient fine-tuning process instead of a data hungry fine-tuning process.

Exploration of the above research question allows us to conclude its capability in our case study. If certain large language models provide good accuracy in our downstream task, then the above-mentioned claim in existing NLP research s correct. If certain multilingual models do not provide good accuracy, then ensemble techniques can provide us with a strong model. In this way, we can utilize the powers of existing transformer based multilingual models for solving challenging NLP application scenarios in a better way. We believe that exploration of the above research question is necessary through our case study. This will help to validate certain claims about novel downstream tasks in a multilingual setting. Depending on the result of this exploration, we would like to come up with a test time ensembling technique which automatically adapts its behaviour based on the presence or absence of the class imbalance problem. Our work will ensure the multilingual NLP researchers to focus on the problem, rather than concentrating on the size of the dataset or presence of class imbalance problem. Real world use cases in multilingual financial domain can be handled fruitfully through our study.

Existing literature present in the financial domain focuses very little on IPO-based NLP applications. To the best of our knowledge, above-mentioned research question has not been studied in any research work present in the financial domain. This paper introduces a credible approach of examining the capability of pretrained multilingual language models through our case study. We also integrate the advantages of these models in proposing our ensembling technique. To summarize, the primary contributions of this study are given below:

- We have created a new multi-lingual, code-mixed corpus called LIC IPO of tweets (text) annotated with sentiment labels (positive, negative and neutral). This will help researchers from varieties of field like corporate finance, behavioural economy, deep learning and NLP.
- We fine-tune various pretrained language models (Bernice [11], TwHIN-BERT [12], MuRIL [13], mBERT [14] and XLM-RoBERTa [15]) to perform sentiment classification over the proposed dataset.

• We have also used our test-time ensemble technique's capabilities to classify multilingual financial sentiments through our proposed dataset.

• Detailed comparative evaluation of our technique over the other models in classifying sentiments of multilingual financial tweets over the proposed dataset has been presented in this work.

• It has been observed that our ensemble technique performs better than most of the multilingual models. Our test-time ensemble approach for multilingual financial tweet analysis during Initial Public Offerings (IPOs) can be a solid research tool to gauge market sentiment and investor interest in a specific company.

The organization of this article is as follows. The preliminary literature to understand multilingual NLP work has been exhibited in Section II. This section also contains a brief survey of research studies related to our work. The details of the creation and annotation of the proposed dataset is described in Section III. The methodology has been explained in Section IV. Experimental evaluations and detailed discussion have been presented in Section V. The conclusion is mentioned in Section VI.

2. Related Work

The number of initial public offerings (IPOs) that were listed between 2021 and 2022 indicates that India's economy is expanding [16]. However, most of the works in corporate finance and business focuses on only IPO underpricing [17]. Aside from that, Yu et al. [18] argued that the market sentiment of financial events is not equivalent to the semantic sentiment of financial news. For specific initial public offerings (IPOs) listed on stock exchanges outside of India, benchmark financial sentiment datasets are available. For example, datasets related to GM's IPO and IPOs listed in Malaysian stock exchange have been discussed in [19] and [20] respectively. The number of benchmark datasets for multilingual financial events such as Indian IPO is very scarce. Our study revolves around identifying investor's opinion during an IPO of certain companies focussing Indian languages and Indian stock exchanges through NLP techniques. In this section, we discuss various research works which are related to our field of study through the following subsections.

2.1. Multilingual NLP literature

Pre-trained language models offer a new experimental platform for tackling a variety of NLP tasks. A large amount of unlabeled data is used to pre-train language models and fine tune them for downstream NLP tasks [21]. In 2018, Jacob et al. [14] proposed BERT (Bidirectional Encoder Representations from Transformers). They demonstrated the superiority of BERT over other deep learning architectures in a variety of NLP tasks such as sentiment analysis, question answering, text generation, text summarization, and so on. The BERT model employs transformer architecture. Transformers have been built using the attention mechanism seen in computer vision literature. Transformers can use differential weights to determine the importance of words in a sentence, which is useful in a variety of NLP tasks. In the work reported by Devlin et al. [22, 14], readers can learn more about the BERT model and its architecture. FinBERT [23] has constructed by finetuning the BERT language model for financial sentiment classification. However, their financial corpus does not include any multilingual data. Few commonly used BERT variants present in multilingual NLP literature are briefly discussed below. However, these models have rarely been investigated for multilingual financial tweets involving Indian IPO in the existing literature.

- **Multilingual BERT (mBERT):** Here, we would like to briefly discuss mBERT [14]; the multilingual variant of BERT. It is pre-trained with 104 languages present across the world. Its performance in a variety of multilingual tasks is promising.
- **XLM-RoBERTa:** XLM-RoBERTa proposed by Alexis et al. [15] is a multilingual model. It is formed by considering previous multilingual models like mBERT and XLM [24]. Their design is also inspired from RoBERTa [25]. They have demonstrated that XLM-RoBERTa outperformed existing multilingual counterparts on a variety of cross-lingual benchmarks.
- **Multilingual Representations for Indian Languages (MuRIL):** MuRIL is developed by Simran et al. [13]. This language model is specifically built for Indian languages. It supports 17 languages. They also claimed that MuRIL outperforms mBERT in various benchmarking tasks. However, they have not compared their results with other multilingual language models present in the literature.
- **IndicBERT:** IndicBERT is a multilingual ALBERT model trained on large-scale corpora. Along with English, IndicBERT [26] focuses on languages from the Indo-Aryan and Dravidian language families. IndicBERT has far fewer parameters than other multilingual language models such as mBERT and XLM-RoBERTa while still providing cutting-edge performance on a variety of NLP tasks.
- **Multilingual models created through tweet data:** Tweets differ from formal text since they are multilingual. Apart from code-mixing, tweets contain *emojis* and *hashtags*. Therefore, people come up with new multilingual models by focusing on Twitter data. In this section, we focus on recently developed multilingual models trained on Twitter data. Following two multilingual models are selected due to its superiority in various NLP tasks.

- **Bernice:** It was introduced by Alexandra et al. [11]. Bernice uses 2.5 billion tweets along with a tweet-focused tokenizer for training purpose. Their training process is inspired by BERTweet [27] and RoBERTa. Bernice was evaluated with multiple multilingual tasks defined over common benchmarking datasets, including sentiment classification. Bernice outperformed XLM-T [28], TwHIN-BERT [12], and various other multilingual models in those tasks. The sentiment classification of multilingual financial tweets has not yet been evaluated through this model.
- **TwHIN-BERT:** Seven billion tweets in more than 100 different languages are used to train TwHIN-BERT [12]. They identify social similarity in the tweets and used it for building their multilingual model. They have used their model for various downstream tasks such as semantic understanding of the text and social recommendation. It is also claimed that this model outperforms other multilingual models in various Tweet benchmark datasets involving tasks such as hashtag prediction and social engagement prediction. However, this model has not been examined with sentiment classification of multilingual financial tweets.

2.2. Twitter-driven studies during IPO

Cornelli et al. [29] discovered that investors' over-optimism can cause IPOs to trade at much higher prices on the day of listing. Despite these evidences, the number of case studies focusing on financial tweets about IPOs is limited [30]. There are very few research works in the literature that focus on Indian IPOs and subsequent analysis of multilingual tweets about those IPOs. Mehta et al. [31] investigated Twitter users' sentiments surrounding Paytm's IPO. However, their dataset is not publicly available. Aside from that, they have limited their research to only tweets in English. They classified the user's tweets using the NVivo tool. Therefore, we want to create a unique, openly available dataset containing multilingual tweets of LIC IPO.

2.3. Sentiment classification of multilingual financial tweets

Financial tweet classification is an active research area [32]. Surprisingly, there are very few works devoted to multilingual tweet classification in the financial domain [33, 34]. We are unable to locate such works in the case of IPO. The work reported by Aysun et al.[35] utilizes the messages of five stocks from the StockTwits platform. There are only two classes (bearish and bullish) in their classification task. The performance of RoBERTa [25] in their classification task is found to be the best. Anil et al.[36] carried out experiments related to the sentiment analysis of Turkish tweets.

They conclude that BERT is better than traditional machine learning algorithms such as logistic regression. Sentiment analysis of financial tweets is carried out by Chen et al. [37] over the FiQA dataset[1]. Their prediction task is divided into two sub-tasks. Firstly, they predict the bullish / bearish sentiment of the tweets. Secondly, they predict the sentiment degree for each tweet. However, they do not consider multilingual financial tweets and BERT models are not employed in their prediction task.

2.4. Ensemble techniques for multilingual sentiment classification in the financial domain

The ensemble learning method brings together multiple weak learners to produce a strong learner. It has been demonstrated that combining transformer-based model predictions performs better in a variety of NLP tasks [38]. It is challenging to discuss every ensemble technique that has been proposed so far. To have a better knowledge of current combiner schemes or ensemble strategies for multi-class classification problems, readers can check into the work done by Sagi et al. [39] and Zhang et al. [40]. Modern test-time ensemble techniques presented in [41, 42] are based on averaging the weights of the fine-tuned models only. In the financial domain, the number of ensemble techniques driven by the transformer based multilingual models is relatively less. In this study, we focus on IPO and majority voting scheme based ensemble techniques driven by various multilingual models. Therefore, it is much more difficult to identify research articles of similar nature.

3. Data Acquisition for LIC IPO Dataset

Data acquisition is the collection of multilingual financial tweets with respect to LIC IPO. This IPO opened on May 04, 2022 and closed on May 09, 2022. Shares were allotted to the retail investors on May 12, 2022. The listing date of this company on Indian stock market was May 17, 2022. Therefore, the time period for data collection is set between May 12, 2022 and May 16, 2022. Within these five days, the financial tweet data about this IPO are collected. This section discusses the experimental protocol used to create the dataset. Necessary description about this LIC IPO dataset is also discussed here.

3.1. Experimental protocol

3.1.1 Data acquisition program

[1] https://sites.google.com/view/fiqa/home?pli=1

Using Tweepy to collect historical Twitter data becomes a standardised process for machine learning practitioners. The data acquisition program in this work is based on Tweepy. A simple tool is created using Tweepy documentation[2] and several GitHub repositories[3]. This tool collects multilingual financial tweets with the hashtag “LIC IPO". A brief summary of the tool creation process is provided below.

- Create a Twitter developer account.
- Obtain the consumer key, consumer secret, access key and access secret from Twitter.
- Install Tweepy using pip.
- Write Python scripts to gather tweets.

3.1.2 Data collection and cleaning

We can use the above tool to download a CSV file containing the extracted multilingual financial tweets during the aforementioned time period. This file contains the date, timestamp, tweet of the user and Twitter user information. Initially, 1000 tweets were collected over the five-day period mentioned above. All of these tweets are associated with the hashtag “LIC IPO”.

In order to guarantee data quality and relevance, the data cleaning procedure for the above-mentioned CSV file entails multiple crucial steps. To prevent redundancy, the data is first loaded and checked for duplicates. If found, they are subsequently eliminated. It is imperative to handle missing values, either through the imputing of missing data or the removal of incomplete entries. Next comes text normalization, which removes special characters and punctuation and converts the text to lowercase. Tokenization divides the text into digestible chunks, and stop word removal gets rid of words that are frequently used but don’t really add much to the meaning. Last but not least, lemmatization or stemming reduces words to their most basic forms, guaranteeing consistency throughout the dataset. Apart from this, we remove special characters, extra spaces, Twitter handles present in the text, irrelevant Twitter metadata. After performing the aforementioned steps on the collected dataset, it is reduced to 960 tweets.

3.1.3 Data annotation

The LIC IPO dataset is a multilingual dataset. With regard to this dataset, there are three sentiment classes: positive, negative, and neutral. Therefore, it is difficult to collect annotators having linguistic proficiency in multiple regional languages used in India. As a result, we use Google translation[4] to translate each multilingual tweet to English, ensuring that annotations for each of those tweets are correct.

Table 1. Languages in the dataset.

Serial number	Language	ISO
1	English	en
2	Hindi	hi
3	Telugu	te
4	Gujrati	gu
5	Marathi	mr
6	Tamil	ta
7	Malayalam	ml
8	Kannada	kn
9	Oriya	or
10	Bengali	bn

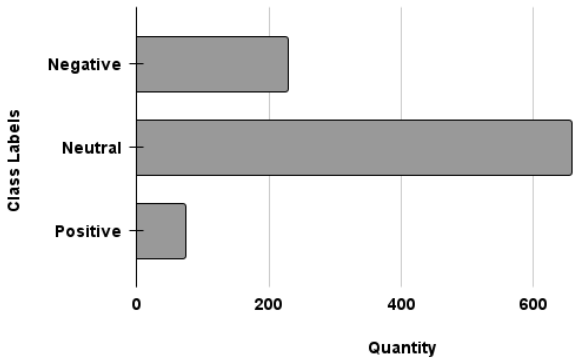

Fig. 1. Classwise statistics of the dataset.

[2] https://docs.tweepy.org/en/latest/getting_started.html
3 https://github.com/keitazoumana/Social-Media-Scraping/blob/main/Twitter%20Scraping.ipynb
4 version: googletrans==3.1.0a0

3.2. Description about the dataset

Figure 1 depicts dataset statistics based on the three sentiment classes discussed previously. In our corpus, negative sentiment labels appear in 229 data points, while positive and neutral sentiment labels appear in 74 and 657 tweets, respectively. The sentiment classes are imbalanced, according to the statistics shown in Figure 1. Due to the nature of our problem statement, there are imbalances in the statistics of the sentiment classes.

Table 2 displays a few examples from our annotated dataset. A total of 4 instances are shown in this table. This shows that the dataset is inherently multilingual, and tweets reflect the code-mixed nature of the corpus. Multilingual tweets in our proposed dataset correspond to 10 different languages. Table 1 is presented to showcase these languages along with their standardized nomenclature (ISO).

Table 2. Few examples from our annotated dataset.

Tweet	Sentiment Class
T1: Millions of Indians investing in the country's biggest listing could turn sour on the equity market if the stock follows the poor performance of its state-run predecessors.	Negative
T2: IPO માં રહી ગયા હોવ તો બે દિવસ રાહ જોઈ લો, આ રીતે સસ્તામાં મળી શકે છે LIC ના શેર Translation: If you have remained in IPO, wait for two days, this way you can get LIC shares cheaply.	Negative
T3: िकस िकस को LIC के शेयर अलॉट हुए है Translation: To whom the shares of LIC have been allotted?	Neutral
T4: LICIPO आज से LIC की GMP बढ़ना शुरू। देखते रहो अब Translation: LICIPO LIC's GMP starts increasing from today. Keep watching now.	Positive

4. Methodology

This section consists of three sections, and the first subsection discusses design of experiments and its rationale. It also consists of the workflow and necessary technical details for understanding our methodology. The second subsection discusses various crucial topics such as data involved in training / testing, experimental setup and fine-tuning process. The third subsection discusses the evaluation metrics used in evaluating our methodology.

4.1. Design of experiments

Two experiments have been designed and we term them as Experiment 1 and Experiment 2 respectively. These are mentioned below and discussed in detail in subsection 4.1.1 and 4.1.2 respectively.

1. Experiment 1: Classification of multilingual financial tweets using five existing multilingual language models over our newly introduced LIC IPO dataset.
2. Experiment 2: Application of our proposed test-time ensemble technique for classification of multilingual financial tweets over the LIC IPO dataset.

4.1.1 Experiment 1

Here, we need to design experiment 1 in such a way that we can successfully classify the multilingual tweets present in our proposed dataset. Five pretrained multilingual models were investigated. These are given below:

- Bernice
- TwHIN-BERT
- MuRIL
- mBERT
- XLM-RoBERTa

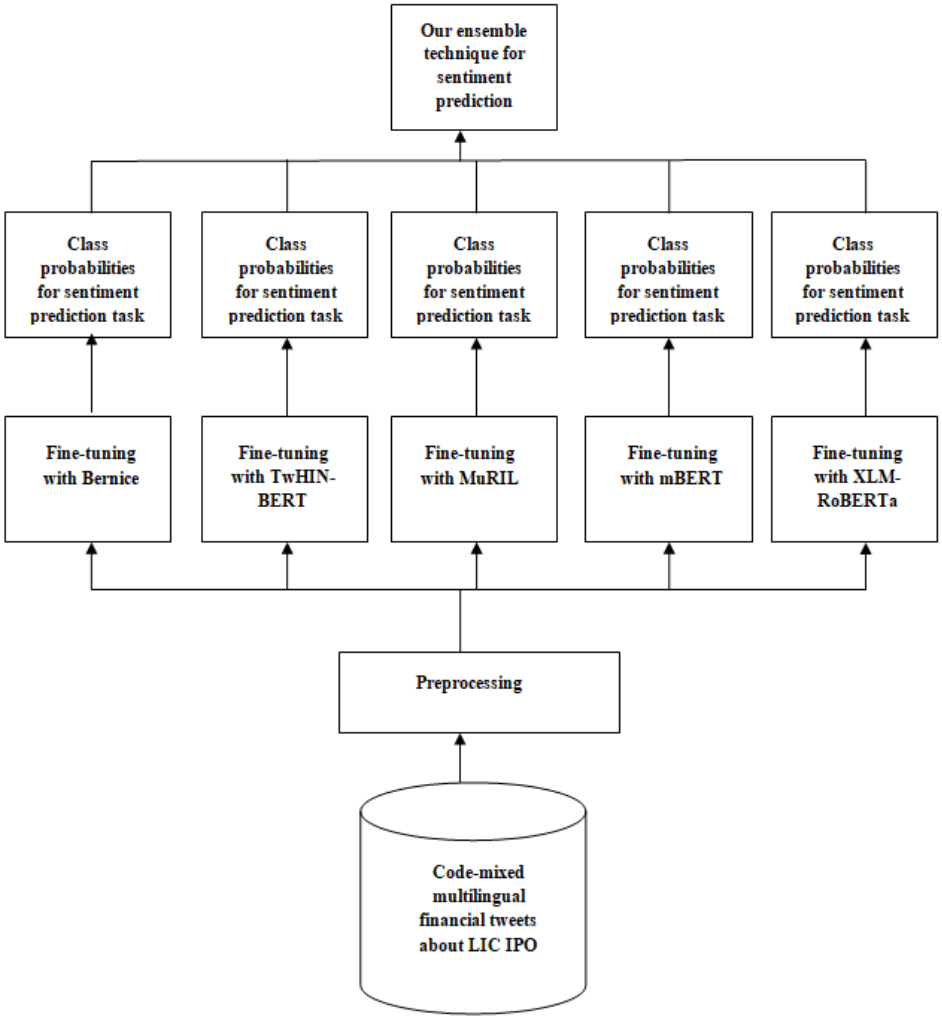

Fig. 2. Workflow for realizing Experiment 1 and Experiment 2

Selection of the above models needs to be discussed here. Bernice is a pre-trained, multilingual encoder created especially for Twitter data [11]. It makes it possible for researchers to glean important insights from complicated and varied Twitter data involving multilingual NLP research [43]. We consider this model for answering our research question and investigating our case study because it is based on Twitter data. In a variety of multilingual social recommendation and semantic understanding tasks, TwHIN-BERT has demonstrated notable gains over other pre-trained models [43]. This demonstrates its robustness and dependability. In designing our experiments, we have selected this model because of such reasons. Due to its specificity for Indian languages, MuRIL has been chosen for our investigation. It supports transliterated text and code-mixed tweets. In a variety of multilingual tasks, this open-source model performs better than others. We chose mBERT for our study because it is an excellent candidate for multilingual NLP research. Owing to a number of features, including benchmark performance across multilingual tasks, the ability to handle code-mixed data in a better way, the capability for transfer learning, and the sizeable community that results from its widespread use. One more multilingual model has been selected in designing our experiments. Due to its extensive use of large-scale data, XLMRoBERTa is well-suited to handle the informal and noisy nature of multilingual social media text. In this preliminary investigation over our proposed dataset, we choose these five language models for experimentation due to two reasons. Firstly, the literature shows that existing multilingual tweets involving Indian languages are mostly processed with these multilingual models. Secondly, although these pretrained models do not require preprocessing, they can easily enable transfer learning capability in different multilingual tasks.

Figure 2 demonstrates the workflow of our experimentation procedure. It will help readers to visualize the experiments carried out over the proposed dataset. In experiment 1, fine-tuning the pre-trained language models will help us in classifying the sentiments of Twitter users from their financial tweets about LIC IPO. In the second experiment, we realize our ensemble technique depicted above. This experiment considers the output produced by the previous experiment i.e. Experiment 1 and predicts the sentiments of financial tweets present in the test set of the LIC IPO dataset.

4.1.2 Experiment 2

Experiment 1 is designed to answer our research question (RQ) and to observe the performance of the state of the art multilingual language models over the LIC IPO dataset. This dataset is having less number of samples along with the class imbalance problem. Experiment 2 is designed while exploring our research question. If outcome of Experiment 1 is not satisfactory, can we build an ensemble technique to lessen the issue of class imbalance and low number of data samples by integrating the power of all those multilingual models? Therefore, ensemble technique development drives the design of Experiment 2. Before providing technical details of this technique, we provide necessary motivation along with key technical recipe for our proposed test-time ensemble technique.

During training, the cost function is crucial for modifying a neural network's weights to produce a more accurate machine-learning model. Traditional loss function applied to a dataset having a class imbalance problem may complicate the scenario [44]. In our scenario, fine-tuning the pretrained multilingual models through traditional loss function may create convergence problem. Using the weighted cross entropy loss function is found to be helpful in training deep learning systems having unbalanced datasets [45].

Most of the people involved in machine learning research are familiar with true positive (TP) rate. When compared to the actual class labels, it assesses how accurately the model predicted the class label. It has been claimed that this TP rate-driven ensemble technique beat various other ensemble techniques such bagging, AdaBoost, Voting etc [46]. This is true for multi-class classification scenarios as well.

The above discussion allows us to come up with a new test-time ensemble technique by integrating the advantages of transformer-based multilingual language models. This technique should not only solve our multi-class classification problem but also helps other people to build a test-time combiner scheme for any type of dataset as well. Our technique will automatically adapt its behaviour depending on whether we are tackling with datasets having class imbalance problem or datasets without having this class imbalance problem. Therefore, our ensemble technique is driven by three things namely: a) TP rate, b) weights used in the above-mentioned loss function and c) rank or position of a class in terms of the number of samples present in the dataset. We introduced Algorithm 1 to realize our proposed transformer-driven ensemble technique. Readers may look into steps 7 and 8 in Algorithm 1 for TP rate, the wight concept is captured in step 5 of Algorithm 1, and step 6 of Algorithm 1 corresponds to rank. This algorithm incorporates our technical intuition and most of the other steps in this algorithm are self-explanatory in nature.

4.2. Model Setup

To evaluate the baseline pretrained language models, it is necessary to discuss the proposed dataset's segregation into training and test sets. In addition, the experimental setup is also discussed in this subsection. Our multilingual NLP task in non-trivial in nature. Therefore, an overview of the fine-tuning process is provided in subsection 4.2.3.

4.2.1 Training and Test data

For the sentiment classification task mentioned in this paper, we use 576 instances in the proposed dataset as the training set. This number corresponds to 60% of the tweets present in the proposed dataset. The remaining 384 instances (40% of the dataset) present in the collected dataset is considered to be the test set. The summary of the training and test set along with the number of instances present in the respective classes are mentioned in table 3.

Algorithm 1: Algorithm for our proposed test-time ensemble technique

---

Input: Evidence, *E* derived after fine-tuning the models (i.e. *C1*, *C2*..).
Output: Predicted sentiment labels for each sample of the test set.
Step 1: Let *N* denotes total number of samples in the test dataset.
Let *M* denotes total number of models (i.e. classifiers) used in the experimentation.
Let *L* denotes total number of classes present in the multi-class classification problem.
Step 2: Create a matrix, *whole* having rows equivalent to *M* and columns equivalent to *N*.
Step 3: Fill the *whole* from evidence, *E* containing predicted class label (p_{ij}) given by model C_i for test sample *j* of the test dataset.

```
for t=1 to t=M do
    for k=1 to k=N do
        whole[t][k] = p_tk
```

Step 4: Create a one dimensional array, result having size equal to *L*. It will contain count values for each class label and initialize all the elements of this array to be 0.
Step 5: Create a one dimensional array, *weight* having size equal to *L*.

for *d=1* to *d=L* **do**

$$\text{weight}[d] = \frac{Number\ Of\ Samples\ Of\ The\ Class_d}{Total\ Number\ Of\ Samples\ Present\ In\ The\ Dataset}$$

Step 6: Create an array *rank* having length equal to *L*. It contains the position of a class in terms of its samples present in the dataset. If total samples of a class (class label 2) is the highest, then position will be 1. This value will be assigned in *rank[2]*.
Step 7: Create a two dimensional array, *TruePositiveRate* having rows equivalent to *M* and columns equivalent to *L*.
Step 8: Fill the *TruePositiveRate* from evidence, *E* containing *true positive rate* of each model for each of the classes present in the multi-class classification problem.
Step 9:

for t=1 to n=N **do**
for k=1 to k=M **do**
for e=1 to e=L **do**
if (*whole[n][k]*==e)
save = k
Considering the remaining values of *whole[t][k]* except the current k and if all of them predict class labels different from current e, execute step 9.1 else 9.2.
Step 9.1
result[e]=result[e]+TruePositiveRate[save][e]+weight[e]*rank[e]
Step 9.2
result[e]=result[e]+TruePositiveRate[save][e]+weight[e]

Step 10: Find out the largest value of *result* array. Let index, v corresponds to the largest value i.e. *result[v]*. This index will be the class predicted by the ensemble technique.

4.2.2 Experimental Setup

We use the PyTorch deep learning framework to fine-tune the five multilingual models mentioned above. Apart from this, Algorithm 1 for the proposed ensemble technique has also been implemented using PyTorch only. Therefore, the workflow (given in figure 2) has been realized by PyTorch. GPU P100 is used for our experimentation. Hugging Face library[5] contains the pre-trained versions of the language models. This library is used in fine-tuning these transformer-based language models. In our experimentation, commonly available Python libraries are also used for dataset preprocessing.

4.2.3 Overview of fine-tuning process

Fine-tuning the pretrained language models is carried out with the help of HuggingFace library. It is a collection of trained machine learning models that are ready for fine-tuning and deployment. A total of four epochs are used for fine-tuning each of the multilingual models. Stochastic gradient descent (SGD) optimizer [47] is used in fine-tuning each of these models. A learning rate of $1e-08$ is used for fine-tuning these models. Various hyperparameters such as learning rate, batch size, number of layers, number of epochs etc. exist in our fine-tuning process. An optimal set of hyperparameters are used in our fine-tuning step after applying a brute force method.

For solving this multi-class classification problem, we are dealing with a dataset having class imbalance issue. Therefore, we utilize the concept of *weight* (as discussed in Algorithm 1) while using the optimizer. Individual PyTorch classes are created for fine-tuning the multilingual language models with the above setting. After training them appropriately, all of them can be evaluated through our test dataset to observe their performance.

Table 3. Summary of the training and test set

Training Set		**Test Set**	
Class label	**Number of tweets**	**Class label**	**Number of tweets**
Negative (class 0)	136	Negative (class 0)	93
Neutral (class 1)	399	Neutral (class 1)	258
Positive (class 2)	41	Positive (class 2)	33

[5] https://huggingface.co/

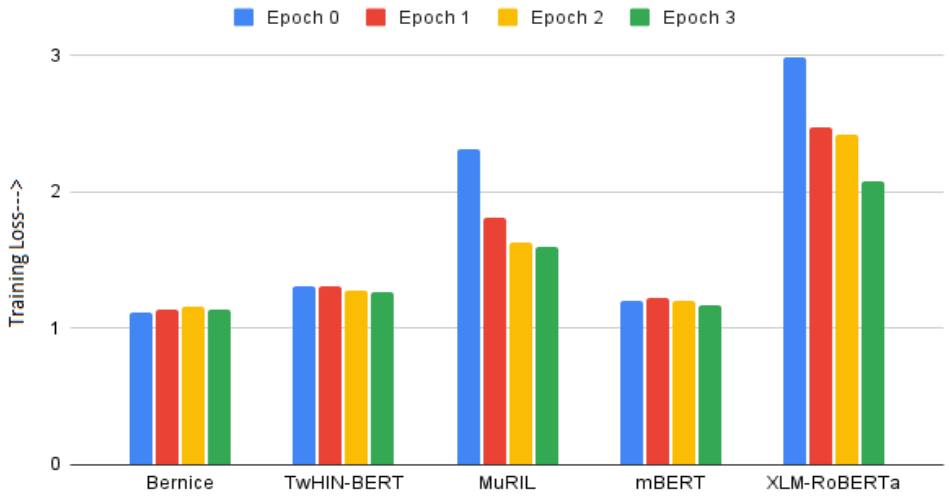

Fig. 3. Training loss over epochs during fine-tuning of various multilingual models

4.3. Evaluation metrics

We have carried out our model evaluation with the help of standard metrics commonly used in the existing deep learning literature. We assess the models' performance using *confusion matrix*, *precision*, *recall*, and *F1-score*. The confusion matrix is a matrix that is used to determine how well five multilingual models / our ensemble technique performs for a given set of test data. Precision is the measure of how accurately the model anticipated the positive tweets to be positive. The positive tweet that was accurately anticipated among all positive tweets is known as recall. The F1-score is the harmonic mean of the precision and recall.

- Precision (P): The formula for precision is given below:

$$\text{Precision} = \frac{\text{True}_{\text{Pos}}}{\text{True}_{\text{Pos}} + \text{False}_{\text{Pos}}} \tag{1}$$

- Recall (R): The formula for recall is given below:

$$\text{Recall} = \frac{\text{True}_{\text{Pos}}}{\text{True}_{\text{Pos}} + \text{False}_{\text{Neg}}} \tag{2}$$

- F1-Score: The formula for F1-Score is given below:

$$\text{F1-Score} = 2 * \frac{P * R}{P + R} \tag{3}$$

True_{Pos}, $\text{False}_{\text{Pos}}$ and $\text{False}_{\text{Neg}}$ means True Positives, False Positives, and False Negatives respectively. These are well-known evaluation metrics used in multi-class classification problems present in machine learning. In Mohammad et al. [48], readers can find out more details about these metrics.

5. Results and Discussion

The experimental results of all the multilingual models and our ensemble technique mentioned in the previous section are presented in this section. All these models were trained using the training set mentioned in Table 3. Readers can see the statistics of the test set which was used to obtain the results. Confusion matrix is used to understand the performance of these models over the proposed dataset. Precision, recall, and F1-score are also used to demonstrate the results.

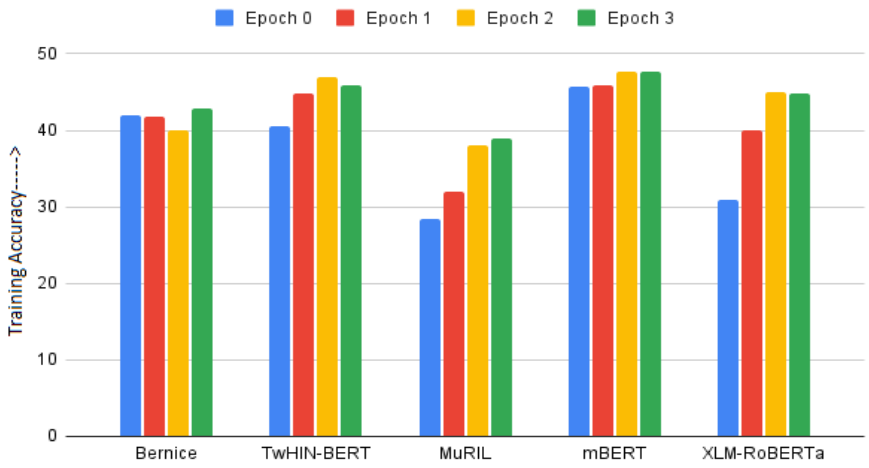

Fig. 4. Training accuracy over epochs during fine-tuning of various multilingual models

5.1. Discussion of output produced by Bernice

For Bernice, the training loss during the fine-tuning process is depicted in figure 3 (group of four bars under the heading Bernice). Similarly, figure 4 shows the training accuracy encountered by Bernice over the four epochs. Figure 5 shows the confusion matrix obtained through Bernice over the test data extracted from LIC IPO dataset. The rows of this matrix show the predicted labels by Bernice. The columns of this confusion matrix show the actual sentiment labels. Considering the negative (class 0) sentiment labels, Bernice can correctly classify 63 multilingual tweets out of 93 multilingual tweets. Considering the neutral (class 1) sentiment labels, Bernice can correctly classify 84 multilingual tweets out of 258 multilingual tweets. For the positive (class 2) sentiment labels, it is found that Bernice can not correctly classify any multilingual tweets.

5.2. Discussion of output produced by TwHIN-BERT

For TwHIN-BERT, the training loss during the fine-tuning process is depicted in figure 3 (group of four bars under the heading TwHIN-BERT). Similarly, figure 4 shows the training accuracy encountered by TwHIN-BERT over the four epochs. Figure 6 depicts the confusion matrix in relation to TwHIN-BERT. The rows of this matrix show TwHIN-BERT's predicted labels. Similarly, the actual sentiment labels are displayed in the columns of this confusion matrix. Considering the negative (class 0) sentiment labels, TwHIN-BERT can correctly classify 16 multilingual tweets. Considering the neutral (class 1) sentiment labels, TwHIN-BERT can correctly classify 193 multilingual tweets out of 258 multilingual tweets. For the positive (class 2) sentiment labels, it is found that TwHIN-BERT can not correctly classify any multilingual tweets.

5.3. Discussion of output produced by MuRIL

For MuRIL, the training loss during the fine-tuning process is depicted in the figure 3 (group of four bars under the heading MuRIL). Similarly, figure 4 shows the training accuracy encountered by MuRIL over the four epochs. Figure 7 depicts the confusion matrix in relation to MuRIL. The rows of this matrix show MuRIL's predicted labels. Similarly, the actual sentiment labels are displayed in the columns of this confusion matrix. Considering the negative (class 0) sentiment labels, MuRIL can correctly classify 21 multilingual tweets. Considering the neutral (class 1) sentiment labels, MuRIL can correctly classify 166 multilingual tweets out of 258 multilingual tweets. For the positive (class 2) sentiment labels, it is found that MuRIL can correctly classify 04 tweets out of 33 multilingual tweets.

5.3. Discussion of output produced by mBERT

For mBERT, the training loss during the fine-tuning process is depicted in the figure 3 (group of four bars under the heading mBERT). Similarly, figure 4 shows the training accuracy encountered by mBERT over the four epochs. Figure 8 shows the confusion matrix obtained through mBERT over the test data extracted from LIC IPO dataset. The rows of this matrix shows the predicted labels by mBERT. The columns of this confusion matrix show the actual sentiment labels. Considering the negative (class 0) sentiment labels, mBERT can correctly classify 40 multilingual tweets. Considering the neutral (class 1) sentiment labels, mBERT can correctly classify 161 multilingual tweets out of 258 multilingual tweets. For the positive (class 2) sentiment labels, it is found that mBERT can not correctly classify any tweets.

5.4. Discussion of output produced by XLM-RoBERTa

For XLM-RoBERTa, the training loss during the fine-tuning process is depicted in the figure 3 (group of four bars under the heading XLM-RoBERTa). Similarly, figure 4 shows the training accuracy encountered by XLM-RoBERTa over the four epochs. Figure 9 depicts the confusion matrix in relation to XLM-RoBERTa. The rows of this matrix

show XLM-RoBERTa's predicted labels. Similarly, the actual sentiment labels are displayed in the columns of this confusion matrix. For the positive (class 0) sentiment labels, it is found that XLM-RoBERTa can not correctly classify any tweets. Considering the neutral (class 1) sentiment labels, XLM-RoBERTa can correctly classify 257 multilingual tweets out of 258 multilingual tweets. For the positive (class 2) sentiment labels, it is found that XLM-RoBERTa can not correctly classify any tweets.

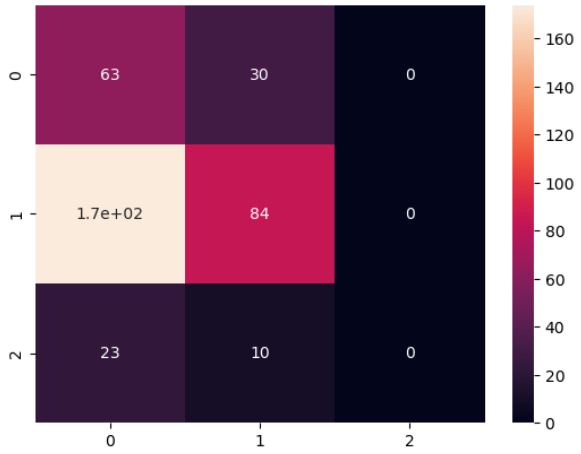

Fig. 5. Confusion matrix obtained through Bernice

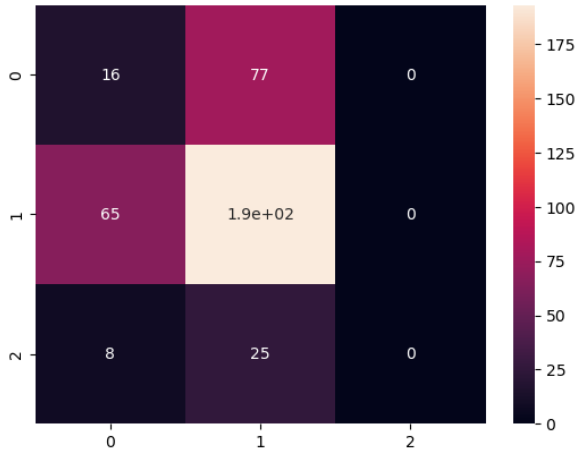

Fig. 6. Confusion matrix obtained through TwHIN-BERT

5.5. Discussion of output produced by our test-time ensemble technique

Figure 10 shows the confusion matrix obtained through our ensemble technique over the test data extracted fromLIC IPO dataset. The rows of this matrix show the predicted labels by our technique. The columns of this confusion matrix show the actual sentiment labels. Considering the negative (class 0) sentiment labels, our technique can correctly classify 16 multilingual tweets. Considering the neutral (class 1) sentiment labels, our technique can correctly classify 203 multilingual tweets out of 258 multilingual tweets. For the positive (class 2) sentiment labels, it is found that our technique can correctly classify 01 tweet out of 33 multilingual tweets. Using different types of transformer-based language models does incur a higher computational cost compared to using a single model. However, the proposed ensemble method has been designed to address the challenges of a complex task where a single model's performance may not be sufficient. The method uses the advantages of multiple language models, and is capable of handling imbalanced datasets and generalizing to different data efficiently. While the performance gain may appear marginal when compared to the most accurate single model in terms of overall accuracy, the ensemble provides a more robust approach to sentiment classification and addresses the class imbalance problem more effectively.

5.6. Summary

Table 3 demonstrates the result produced by these five models and our ensemble technique in terms of precision, recall and F1-score. The overall accuracy of Bernice in this multi-class classification problem is 38.28%. Under the same experimental conditions, the accuracy of TwHIN-BERT in the same test set is only 54.42%. Similarly, accuracies of MuRIL, mBERT and XLM-RoBERTa are 49.73%, 52.34%, and 66.92% respectively. From figure 4, it is evident that the training loss diminishes with the passage of each epoch for MuRIL and XLM-RoBERTa. From figure 5, the

training accuracy improves as epochs pass for these two models. Apart from these two models, TwHIN-BERT also mildly follows a similar kind of pattern. However, Bernice and mBERT do not follow this pattern. From table 3, it is observed that XLM-RoBERTa produces the highest accuracy in predicting the sentiment labels from the test set of LIC IPO dataset. However, this model should not be used because of its performance over negative and positive class labels. This model is biased towards the class label having the maximum number of samples. Apart from this, TwHIN-BERT performs better as compared to the remaining individual multilingual models. However, this model is not suitable for predicting positive sentiment labels.

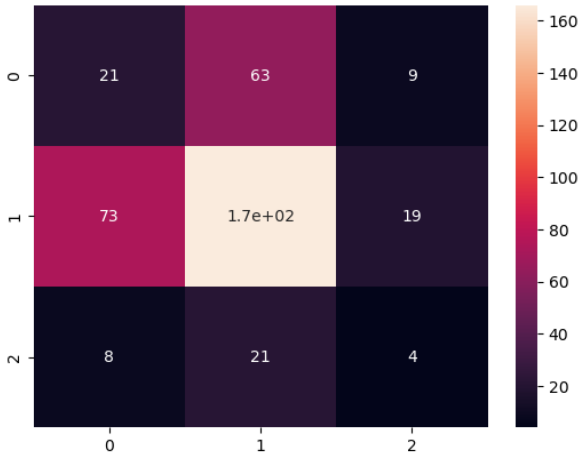

Fig. 7. Confusion matrix obtained through MuRIL

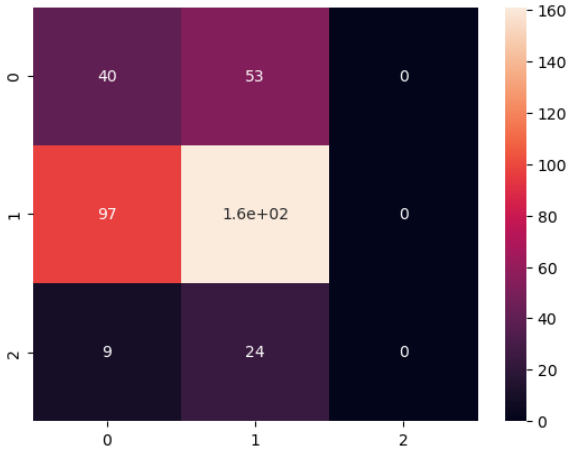

Fig. 8. Confusion matrix obtained through mBERT

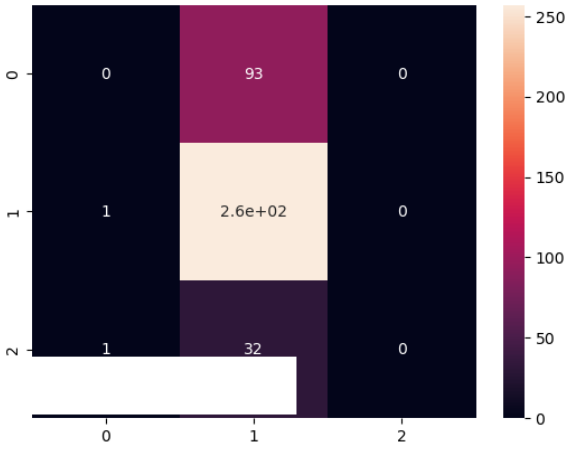

Fig. 9. Confusion matrix obtained through XLM-RoBERTa

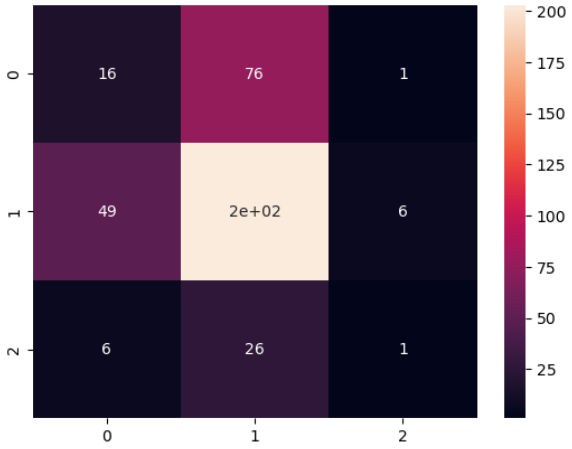

Fig. 10. Confusion matrix obtained through our ensemble technique

Table 4. Experimental results of various techniques over the test set of LIC IPO dataset

Evaluation Technique	Class labels	Precision	Recall	F1-Score	Accuracy
Bernice	Negative	0.24230769	0.67741935	0.35694051	38.28
	Neutral	0.67741935	0.3255814	0.43979058	
	Positive	0	0	0	
TwHIN-BERT	Negative	0.17977528	0.17204301	0.17582418	54.42
	Neutral	0.65423729	0.74806202	0.69801085	
	Positive	0	0	0	
MuRIL	Negative	0.20588235	0.22580645	0.21538462	49.73
	Neutral	0.664	0.64341085	0.65354331	
	Positive	0.125	0.12121212	0.12307692	
mBERT	Negative	0.2739726	0.43010753	0.33472803	52.34
	Neutral	0.67647059	0.62403101	0.64919355	
	Positive	0	0	0	
XLM-RoBERTa	Negative	0	0	0	66.92
	Neutral	0.67277487	0.99612403	0.803125	
	Positive	0	0	0	
Ours	Negative	0.22535211	0.17204301	0.19512195	57.29
	Neutral	0.66557377	0.78682171	0.72113677	
	Positive	0.125	0.03030303	0.04878049	

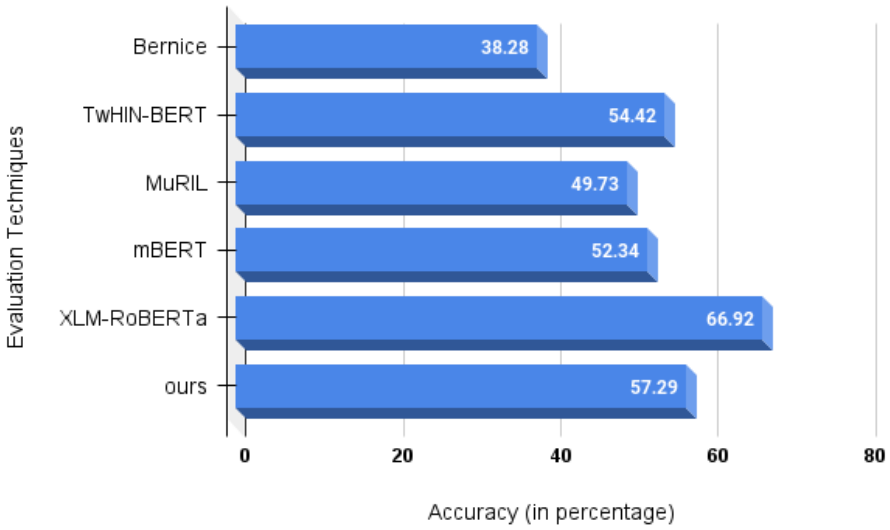

Fig. 11. Accuracy of various language models and our test-time ensemble technique over the test of LIC IPO dataset

Above discussion leads to an interesting observation which is linked to our research question (RQ). It has been found that fine-tuning the multilingual language models to a novel downstream task is not going to provide excellent accuracy on the test dataset. This is supported by our experimental evaluation mentioned in Table 3. Although these pre-trained models have experienced lots of textual data in different languages; it is not able to perform well in our IPO-driven multilingual sentiment classification task. Through our case study, we can conclude that these models have not

encountered a large amount of code-mixed, multilingual financial data during the fine-tuning process. Therefore, fine-tuning these models on less amount of data does not help these models to generalize to unseen financial tweets involving this IPO. The attention mechanism within transformers need to be improved in order to learn from fewer examples to compete with human beings' capability of learning from very few examples. Therefore, the multilingual code-mixed corpus presented in this paper may encourage NLP researchers to address the shortcomings of transformer-based multilingual language models.

Our ensemble technique produces an accuracy of 57.29% over the same test set of LIC IPO (see figure 12). Considering all these facts along with the data in table 3, it can be concluded that our test-time ensemble technique is better in classifying sentiments of multilingual tweets present in the test set of LIC IPO dataset. This would not have been possible if we don't explore our research question systematically. Our ensemble technique exploits the advantages of transformer-based multilingual models. Algorithm 1 considers several technical ingredients to make this ensemble technique a credible one. Algorithm 1 is designed to leverage the diversity of the base models (transformer-based language models used in our scenario) even if their errors are not completely uncorrelated. This diversity is further supported by the observation that different models exhibit varying performance across different classes in the financial sentiment analysis task. Instead of just averaging [41, 42] or using a simple majority vote, our algorithm (algorithm 1) is more sophisticated. Algorithm 1's mechanism for weighting and combining predictions further improves performance, even if there are correlations in the errors. The empirical results indicate that the ensemble is more effective than most of the individual models in the given task. Further research could be done to explicitly quantify error correlation.

In addition to testing current language models, we take into account other pertinent metrics like precision, recall, and F1 score when assessing our approach. Since these metrics are frequently used in deep learning and machine learning research, readers can look into table 3 and interpret the results on their own. Although we focus on accuracy in this summary, this metric is not the only one that matters. We agree that it is important to understand how and why a language model predicts certain things, particularly when it comes to sensitive financial applications. Part of the ongoing research involves analysing the interpretability of our ensemble method and the language models. Data augmentation is a commonly used method for handling class imbalance problem. However, our algorithm for the proposed ensemble method addresses fewer amounts of data along with class imbalance problem in a unified manner without the necessity of data augmentation. Testing our ensemble method in other similar use cases in other application domain is also a part of the ongoing research.

5.7. Ethical considerations

The ethical guidelines are followed in conducting this research [49]. No personal information about Tweeter users is disclosed by our proposed dataset. When pre-processing our dataset, we complied with data privacy regulations. Despite the fact that this research study is centered on financial event data, it does not recommend any NLP technique that involves the reader's financial gain or loss. In our case study, we use only multilingual NLP models. Therefore, certain biases may be present in those original models. Reporting all ethical dimensions of our proposed ensemble method is part of ongoing research.

6. Conclusion

Multilingual financial sentiment analysis during a financial event is essential for research works involving NLP, financial media studies and corporate finance. In this work, we focus on a financial event i.e. LIC IPO. Our study has explored the fact that there is a lack of proper experimental data and awareness of the true power of transformer-based multilingual models in financial sentiment analysis in the literature. We have introduced LIC IPO dataset in this work which is challenging in nature. This is because; it is multilingual and code-mixed in nature. Apart from this, class imbalance problem having fewer amounts of data is also a characteristic of this dataset. We have investigated five multilingual models for sentiment analysis of multilingual financial tweets through our LIC IPO dataset. We report their performance over the test set to understand the true capabilities of them in our novel downstream task. This work also puts forward a test-time ensemble technique by considering these language models. Our ensemble technique performs better than most of the multilingual models in this multilingual financial tweet classification task. Our research will open up new research avenues in several IPO based multilingual NLP research work and financial communication. This technique is proposed for handling fewer amounts of test data and datasets having class imbalance problem. Extensive evaluation of our ensemble approach in different scenarios is part of ongoing research.

Acknowledgement

We would like to thank A. Azam for his assistance with data collection. We are grateful to Dr. S. Lonka and Dr. A. Bhowmick for their insightful discussions in financial data analysis using deep learning.

References

[1] Tuan Hao Ly and Khanh Nguyen. Do words matter: Predicting ipo performance from prospectus sentiment. In 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pages 307–310. IEEE, 2020.

[2] Elena Fedorova, Sergei Druchok, and Pavel Drogovoz. Impact of news sentiment and topics on ipo underpricing: Us evidence. International Journal of Accounting & Information Management, 2021.

[3] Jim Kyung-Soo Liew and Garrett Zhengyuan Wang. Twitter sentiment and ipo performance: A cross-sectional examination. Journal of Portfolio Management, 42(4):129, 2016.

[4] Erkut Memi¸s, Hilal Akarkamçı, Mustafa Yeniad, Javad Rahebi, and Jose Manuel Lopez-Guede. Comparative study for sentiment analysis of financial tweets with deep learning methods. Applied Sciences, 14(2):588, 2024.

[5] Carol Myers-Scotton. Duelling languages: Grammatical structure in codeswitching. Oxford University Press, 1997.

[6] Hatem Haddad, Ahmed Cheikh Rouhou, Abir Messaoudi, Abir Korched, Chayma Fourati, Amel Sellami, Moez Ben HajHmida, and Faten Ghriss. Tunbert: Pretraining BERT for tunisian dialect understanding. SN Comput. Sci., 4(2):194, 2023.

[7] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. Science, 350(6266):1332–1338, 2015.

[8] Rupam Bhattacharyya, Zubin Bhuyan, and Shyamanta M Hazarika. Inferring semantic object affordances from videos. In Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part III 5, pages 278–290. Springer, 2021.

[9] Tom B Brown. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

[10] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682, 2022.

[11] Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. Bernice: A multilingual pre-trained encoder for twitter. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6191–6205, 2022.

[12] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. arXiv preprintarXiv:2209.07562, 2022.

[13] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. Muril: Multilingual representations for indian languages. arXiv preprint arXiv:2103.10730, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116, 2019.

[16] Gopal Acharya and Dushyant Mahadik. Underpricing effect in initial public offerings: A systematic review and bibliometrics analysis. Available at SSRN 4541672, 2023.

[17] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. Financial sentiment analysis: Techniques and applications. ACM Computing Surveys, 56(9):1–42, 2024.

[18] Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. Multi-source aggregated classification for stock price movement prediction. Information Fusion, 91:515–528, 2023.

[19] Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. Gold-standard for topic-specific sentiment analysis of economic texts. In LREC, volume 2014, pages 2152–2157, 2014.

[20] Ali Albada, Muataz Salam Al-Daweri, Rabie A Ramadan, Khalid Al Qatiti, Li Haoyang, and Peng Shutong. Determinates of investor opinion gap around ipos: A machine learning approach. Intelligent Systems with Applications, page 200420, 2024.

[21] Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. Improving bert fine-tuning via self-ensemble and selfdistillation. arXiv preprint arXiv:2002.10345, 2020.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019.

[23] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. arxiv 2019. arXiv preprint arXiv:1908.10063, 2019.

[24] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. Advances in neural information processing systems, 32, 2019.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[26] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In Findings of EMNLP, 2020.

[27] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200, 2020.

[28] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 258–266, 2022.

[29] Francesca Cornelli, David Goldreich, and Alexander Ljungqvist. Investor sentiment and pre-ipo markets. The journal of finance, 61(3):1187–1216, 2006.

[30] Luca Gregori Gian, Camilla Mazzoli Luca, Marinelli, and Severini Sabrina. The social side of ipos: Twitter sentiment and investorsâC™ attention in the ipo primary market. African Journal of Business Management, 14(12):529–539, 2020.
[31] Meera Mehta, Shivani Arora, Shikha Gupta, Arun Jhulka, et al. Social listening through sentiment analysis of twitter data: a case study of paytm ipo. SocioEconomic Challenges, 6(3):39–47, 2022.
[32] Weiwei Jiang. Applications of deep learning in stock market prediction: recent progress. Expert Systems with Applications, 184:115537, 2021.
[33] Ross Gruetzemacher and David Paradice. Deep transfer learning & beyond: Transformer language models in information systems research. ACM Computing Surveys (CSUR), 54(10s):1–35, 2022.
[34] Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. Transformer-based deep learning models for the sentiment analysis of social media data. Array, page 100157, 2022.
[35] Aysun Bozanta, Sabrina Angco, Mucahit Cevik, and Ayse Basar. Sentiment analysis of stocktwits using transformer models. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 1253–1258. IEEE, 2021.
[36] Zekeriya Anil Guven. Comparison of bert models and machine learning methods for sentiment analysis on turkish tweets. In 2021 6th International Conference on Computer Science and Engineering (UBMK), pages 98–101. IEEE, 2021.
[37] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Fine-grained analysis of financial tweets. In Companion Proceedings of the The Web Conference 2018, pages 1943–1949, 2018.
[38] Aapo Pietil ̈anen and Shaoxiong Ji. Aaltonlp at semeval-2022 task 11: Ensembling task-adaptive pretrained transformers for multilingual complex ner. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1477–1482, 2022.
[39] Omer Sagi and Lior Rokach. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1249, 2018.
[40] Hongzhi Zhang and M Omair Shafiq. Survey of transformers and towards ensemble learning using transformers for natural language processing. Journal of big Data, 11(1):25, 2024.
[41] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. Advances in Neural Information Processing Systems, 35:10821–10836, 2022.
[42] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7959–7971, 2022.
[43] George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, and Dimosthenis Kyriazis. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Computing and Applications, 35(29):21415–21431, 2023.
[44] Ziyun Zhou, Hong Huang, and Binhao Fang. Application of weighted cross-entropy loss function in intrusion detection. Journal of Computer and Communications, 9(11):1–21, 2021.
[45] Yaoshiang Ho and Samuel Wookey. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE access, 8:4806–4813, 2019.
[46] Artittayapron Rojarath and Wararat Songpan. Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. Applied Intelligence, 51:4908–4932, 2021.
[47] L éon Bottou. Stochastic gradient descent tricks. Neural Networks: Tricks of the Trade: Second Edition, pages 421–436, 2012.
[48] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2):1, 2015.
[49] Radu Stefan, George Carutasu, and Marian Mocan. Ethical considerations in the implementation and usage of large language models.

Authors' Profiles

R. Bhattacharyya completed his masters in Information Technology (M.Tech in IT) from the Tezpur University, India in 2014. He obtained his Ph.D. in the year 2019 from Tezpur University, India. After five long years as the faculty member of IIIT Bhagalpur, he joined Gauhati University, India. During his tenure at IIIT Bhagalpur, he also established Computer Programming laboratory. His research interests include knowledge representation, applied AI, cognitive vision, simulated robotics and applications of deep learning. He was a recipient of meritorious scholarship by University of Stirling, Scotland, UK, for attending and delivering a talk on International IEEE/EPSRC Workshop on Autonomous Cognitive Robotics (COGROB) in the year 2014. He has also been selected for the Young Faculty Research Grant by Gauhati University for the year 2023-2024. He published research papers in reputed international journals and conferences. He serves as a reviewer in various reputed international journals and conferences.

How to cite this paper: Rupam Bhattacharyya, "Classification of Multilingual Financial Tweets Using an Ensemble Approach Driven by Transformers", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.17, No.2, pp. 51-67, 2025. DOI:10.5815/ijieeb.2025.02.02