Information Processing and Management 59 (2022) 102756

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

An ensemble model for classifying idioms and literal texts using BERT and RoBERTa

J Briskilal [*], C.N. Subalalitha

SRM Institute of Science and Technology, Potheri, Kattankulathur, Chengalpattu, Tamilnadu, India

ARTICLE INFO

Keywords:
BERT
RoBERTa
Ensemble model
Idiom
Literal classification

ABSTRACT

An idiom is a common phrase that means something other than its literal meaning. Detecting idioms automatically is a serious challenge in natural language processing (NLP) domain applications like information retrieval (IR), machine translation and chatbot. Automatic detection of Idioms plays an important role in all these applications. A fundamental NLP task is text classification, which categorizes text into structured categories known as text labeling or categorization. This paper deals with idiom identification as a text classification task. Pre-trained deep learning models have been used for several text classification tasks; though models like BERT and RoBERTa have not been exclusively used for idiom and literal classification. We propose a predictive ensemble model to classify idioms and literals using BERT and RoBERTa, fine-tuned with the TroFi dataset. The model is tested with a newly created in house dataset of idioms and literal expressions, numbering 1470 in all, and annotated by domain experts. Our model outperforms the baseline models in terms of the metrics considered, such as F-score and accuracy, with a 2% improvement in accuracy.

1. Introduction

An idiom is a commonly used expression wherein the real meaning differs from what is conveyed. Example 1, "*let's paint the town red*", is an idiomatic expression that means let's have a great time. Likewise, the idiomatic expression in Example 2, "*she has a bun in the oven*" means she's pregnant. The literal meaning of a word or phrases is their most evident or non-figurative sense. Example 3, "*The striker shot the ball right into the arms of the keeper*" is a Literal expression. Disambiguating idioms and literal expressions may be deemed a classification task. Natural language processing (NLP) applications such as parsing, semantic tagging, machine translation (MT), document indexing and query processing in information retrieval (IR) and question answering systems achieve high accuracy when idiom detection and classification are incorporated (Verma & Vuppuluri, 2015).In an automatic MT system, for instance, idiom recognition is critical to generating meaningful translations. In example 4, the idiomatic sentence, "*I feel under the weather*", is quite literally translated as"நான்வானிலைக்குகீழ்உணர்கிறேன்", though the input idiomatic sentence means "I feel ill", and the translation must rightly read "எனக்குஉடல்நிலைசரியில்லை". In this paper, we attempt to classify idioms and literal expressions using ensemble techniques that combine two deep learning models, BERT (Devlin, Chang, Lee & Toutanova, 2018) and RoBERTa (Liu et al., 2019).

Rule-based generalization is used in idiom recognition, and context-based classification to classify idioms and literal sentences

[*] Corresponding author.
E-mail addresses: briskilj@srmist.edu.in (J. Briskilal), subalaln@srmist.edu.in (C.N. Subalalitha).

https://doi.org/10.1016/j.ipm.2021.102756
Received 19 April 2021; Received in revised form 25 August 2021; Accepted 5 September 2021
Available online 26 September 2021

(Shaikh, 2020). More recently, crowdsourcing has been utilized to detect the sentiment annotations of idiomatic expressions. In all, 5000 frequently occurring idioms were identified with this method (Jochim et al., 2018). Numerous methodologies have attempted to classify idioms and literals, though none employed ensemble pre-trained models like BERT and RoBERTa. This paper attempts to classify idiomatic and literal sentences, using an ensemble method, with good accuracy.

In sum, the main contributions of this paper are twofold:

1. Proposing an ensemble model to classify idioms and literal expressions using BERT and RoBERTa and
2. Constructing a dataset comprising 1470 idiomatic and literal sentences annotated by domain experts.

The rest of the paper is organized as follows: Section 2 discusses the background to the study, and Section 3 presents a literature survey. Section 4 describes the proposed ensemble model and experimental setup for the classification of idioms and literal expressions, while Section 5 concludes the paper and offers suggestions for future work.

2. Background

2.1. BERT

Bidirectional Encoder Representations from Transformers (BERT) is a new language representation model released by Google (Devlin et al., 2018). BERT's purpose is to pre-train deep bidirectional representations, from unlabeled text in every layer, by jointly configuring the context to the left and right of a given text. BERT can be fine-tuned by adding an extra output layer to build a state-of-the-art model to undertake NLP tasks like question answering and classification.

BERT works in two steps, pre-training and fine-tuning. The BERT model is built, based on the Transformer encoder (Vaswani et al., 2017). There are two models, BERT base and BERT large, and their sizes are given below:

Bert base ($L = 12$, $H = 768$, $A = 12$, Total Parameters=110 M)

Bert large ($L = 24$, $H = 1024$, $A = 16$, Total Parameters=340 M)

Where L denotes the number of layers (Transformer blocks), H the hidden size, and A the number of self-attention heads.

In this paper, we take the BERT base model as the ensemble technique to classify idiomatic and literal sentences, using Word Piece embedding with a 30,000-token vocabulary. BERT has special tokens for separation [SEP], classification [CLS] and padding [PAD], and an unknown [UNK] token as well.

The BERT model is trained with two different tasks, the masked language model (MLM) and next sentence prediction (NSP). In order to train the deep bidirectional model in the MLM, input tokens are randomly chosen and masked, following which predictions are made for the masked tokens.

In MLM, In order to train a deep bidirectional model, some input tokens are chosen randomly In the sentence, "*That's what she said*", in Example 5, the tokens 'what' and 'said' are randomly masked as That's [MASK] and she [MASK].

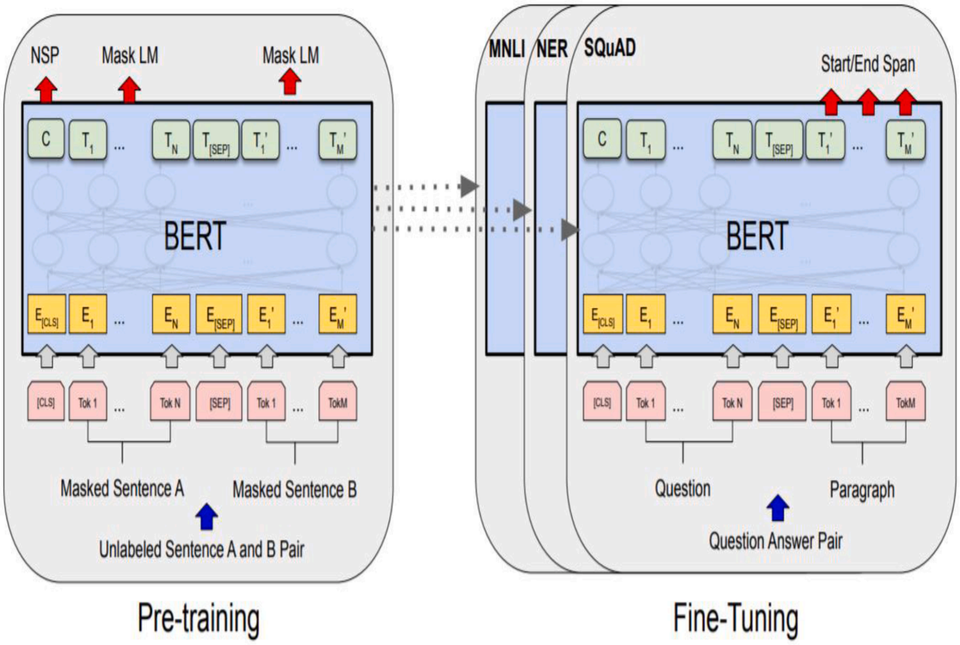


Fig. 1. Pre-training and Fine-tuning methods of BERT.

The NSP technique helps establish the relationship between sentences in the model. For instance, to ascertain if there exists a relationship between two sentences, A and B, the NSP checks if 50% of the pre-trained example consists of A followed by B.

If A and B are found adjacent to each other 50% of the time, they are labeled IsNext, or else the label NotNext is assigned to them. The MLM and NSP are illustrated in Example 5. Fig. 1

Input = [CLS] That's [MASK] she [MASK]. [SEP] Hahaha, nice! [SEP]
Label = IsNext
Input = [CLS] That's [MASK] she [MASK].oh! you ignorant [MASK]! [SEP]
Label = NotNext.
Fig. 2. Shows the input representations of Bert (Devlin et al., 2018).

2.2. RoBERTa

Robustly Optimized BERT Pre-training Approach (RoBERTa) (Liu et al., 2019) is an extension of Bert model. Issues in the BERT model were identified by Facebook AI Research (FAIR) and an optimized, robust version of BERT was built.

RoBERTa model is trained with bigger batches and longer sequences. By giving large batch sizes improve end-task accuracy compared to BERT.

A dynamic masking pattern introduced in RoBERTa duplicates the training data and carries out a range of masking strategies.

When data is passed into the RoBERTa model, different masking strategies are performed every single time. The BERT model, on the other hand, uses a static masking strategy and undertakes masking only during data preprocessing.

Next section describes the Literature survey.

3. Literature survey

This section discusses state-of-the-art approaches applicable to the proposed work, which focuses chiefly on idiom classification as a text classification task, using an ensemble model.

A literature survey has been carried out from four perspectives: the first and second analyze existing text classification work that uses the BERT and RoBERTa models, respectively, while the third and fourth explore work on idioms and ensemble models, respectively.

3.1. Text classification works using BERT model

Gonzalez et al. (2020) undertook text classification for four different datasets (IMDB, RealorNot, Portuguese news and Chinese hotel reviews) using the BERT model. The IMDB dataset was examined to classify 5000 movie reviews. The RealorNot tweet dataset was used for binary classification, principally to ascertain whether or not texts were tweets, and categorize them accordingly. A Portuguese news dataset and a Chinese hotel review dataset were studied to classify news and reviews, respectively. This paper compared the BERT approach with traditional machine learning techniques (González-Carvajal & Garrido-Merchán, 2020).Ashutosh et al. (2019) proposed using the BERT model and a knowledge distillation approach for document classification to improve the efficiency of baseline LSTM variants.Their model was tested on four different datasets, Reuters-21578, the AAPD academic paper dataset, IMDB and Yelp '14. The Reuters-21578 and AAPD are multi-label datasets (Adhikari, Ram, Tang & Lin, 2019).Jihang et al. (2019) examined the classification of events in Spanish text using a Uruguayan newspaper, based on its factuality status. The BERT multilingual-cased model was utilized for multilingual text classification, and a FACT dataset for classification (Mao & Liu, 2019).Sun et al. (2019) proposed an array of fine-tuning methods to provide a general solution for fine-tuning a BERT model to perform different types of text classification, including that of sentiments, questions and topics.

The IMDB movie review and Yelp review datasets were used for sentiment classification, The open-domain, fact-based TREC dataset was used for question classification, and the large-scale AG News dataset for topic classification (Sun, Qiu, Xu & Huang, 2019). Issa (2020) detected humor in short texts using BERT sentence embedding, and the proposed model outperformed baseline models like the RNN. The ColBERT dataset consists of 200k labeled humor and non-humor texts (Annamoradnejad & Zoghi, 2020).Manish et al.

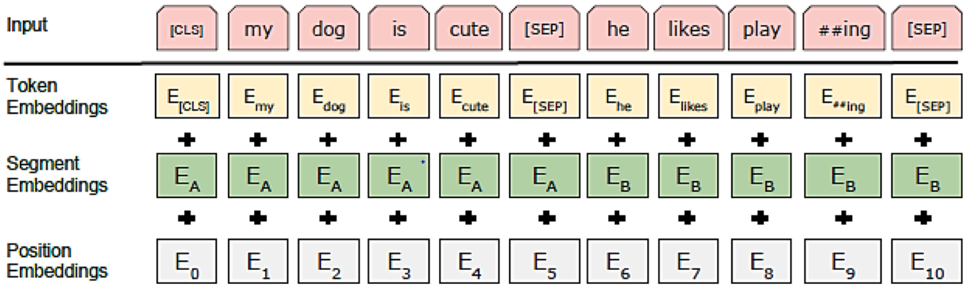


Fig. 2. Input Representation of BERT.

(2019) undertook sentiment classification using the BERT model on the Stanford Sentiment Treebank (SST) dataset (Munikar, Shakya & Shrestha, 2019).

3.2. Text classification works using RoBERTa model

Arup et al. (2020) discussed aggression detection-based classification using the RoBERTa and SVM classifiers. An aggression-identification dataset was shared in three languages – English, Hindi and Bangla. The RoBERTa model used for the English sub-tasks outperformed the SVM classifier with an F-score of 80% (Baruah, Das, Barbhuiya & Dey, 2020).Wexiong et al. (2020) studied aspect category sentiment analysis using the RoBERTa model. A fine-grained sentiment analysis public dataset from AI Challenger was used for classification (Liao, Zeng, Yin & Wei, 2021).Public dataset (Fine grained sentiment analysis) from AI challenger has been used for classification purpose (Liao et al., 2021). Ankit et al. (2020) detected and classified mental illness posts on social media using RoBERTa. In all, 3000 posts were taken from subreddits and used as a dataset for five class labels. RoBERTa outperformed the BERT and LSTM models (Murarka, Radhakrishnan & Ravichandran, 2020).

3.3. Works on classification of idioms

Peng et al. (2018) proposed an algorithm to automatically classify idiomatic and literal expressions. A bag-of-words topic representation was applied to paragraphs to classify them as idioms or literal expressions, based on the topics in question. Four datasets were used in this approach for classification, including BlowWhistle (a total of 78 examples, with 27 idioms and 51 literals), MakeScene (a total of 50 examples, with 30 idioms and 20 literals), LoseHead (a total of 40 examples, with 21 idioms and 19 literals) and TakeHeart (a total of 81 examples, with 61 idioms and 20 literals) (Peng, Feldman & Vylomova, 2018). Irena et al. (2017) proposed an automated approach to enrich sentiment analysis using idiom-based features. A set of rules to recognize idioms and sentiment polarities to execute classification was included, and a gold standard dataset used for the purpose (Spasić, Williams & Buerki, 2017). Naziya (2020) used rule-based generalization and context-based classification to classify idioms and literals in paragraphs (Shaikh, 2020) Liu et al. (2018) proposed a neural network-based approach to recommend the use of idioms in writing essays. Similarity was calculated between a given context and candidate idiom recommendations (Liu, Liu, Shan & Wang, 2018).

Xianyang et al. (2020) carried out token-level metaphor classification on the VUA (VU Amsterdam Metaphor corpus) and TOEFL corpus using a BERT model (Chen, Leong, Flor & Klebanov, 2020). Liu et al. (2017) proposed a supervised ensemble model to classify idioms and literal expressions, using late and early fusion methods (Liu & Hwa, 2017). Charles et al. (2018) proposed an approach to collect the sentiment annotations of idiomatic expressions using crowdsourcing. The Sentiment Lexicon of Idiomatic Expressions (SLIDE), which thus came into being, is much larger than previous lexicons (Jochim et al., 2018). Mona et al. (2009) proposed a supervised learning approach to classify multiword expressions from literals using the VNC-Tokens (verb-noun construction) dataset (Diab & Bhutada, 2009).

3.4. Text classification works using ensemble models

Julian et al. (2020) proposed a bagging-based ensemble model to classify 6000 aggressive and offensive social posts using multiple fine-tuned BERT models (Risch & Krestel, 2020). Tadej et al. (2020) advocated a contextual embedding-based approach to classify idioms and literal expressions using an ensemble BERT and ELMo model. A Slovene dataset was used for the monolingual classification (Škvorc, Gantar & Robnik-Šikonja, 2020). Tanvi et al. (2020) classified conversational social media text using an ensemble RoBERTa and ALBERT model to achieve a higher F-score of 3% more than that of the baseline models. The Get it #offMyChest dataset used for classification constitutes assorted conversations on social media (Dadu, Pant & Mamidi, 2020).Yandrapati et al. (2020) classified 20 K Covid-related tweets using an ensemble CT-BERT and RoBERTa-based approach. Their model achieved a higher F-score than that of the baseline model (Babu & Eswari, 2020). Ting et al. (2019) proposed an ensemble BERT and BiLSTM model to differentiate between fake and genuine news and classify them accordingly. The 321k news titles created for the WSDM Challenge 2019 were used for classification (Su, Macdonald & Ounis, 2019). Verena et al. (2020) attempted to detect propaganda techniques in news articles using a BERT and BiLSTM-based ensemble approach. Articles pulled from 50 news outlets were annotated and used for the purpose (Blaschke, Korniyenko & Tureski, 2020).

The existing research makes it clear that no ensemble methodologies were used to classify idioms and literal texts. Machine learning ensemble models incorporate decisions from multiple models to enhance overall accuracy.

Ensemble models are meta-algorithms that combine several machine learning and deep learning classifiers into a predictive model for reduced variance, bias and improved accuracy. Noise, bias and variance are major causes for errors in learning models. Ensemble methods, which minimize these factors, are designed to enhance the stability and accuracy of machine learning algorithms (Opitz & Maclin, 1999).

In this paper, we attempt to classify idioms and literals using an ensemble BERT-RoBERTa combine. The BERT and RoBERTa models are fine-tuned by training them with the TroFi dataset. To this end, furthermore, we have created a dataset of our own, comprising 1470 idioms and literal expressions, annotated by domain experts and evaluated using an inter-annotator agreement.

Next section describes the proposed experimental setup for classification of idioms and literals.

4. Proposed ensemble model and experimental setup for idioms and literals classification

Fig 3. Shows the proposed ensemble model architecture. The TroFi Metaphor dataset from the Wall Street Journal Corpus Release 1, with 3737 idiomatic sentences and literal phrases (Birke & Sarkar, 2006), is taken and used to train the BERT and RoBERTa models. The predictions of the two models are combined and the weighted average method used to predict the final class labels, namely, idioms and literals. The model is also tested with our in house dataset. We have constructed this dataset by extracting various idiomatic sentences from different websites and equivalent literals sentences were framed and finally it was annotated by domain experts, comprising 1470 idiomatic sentences and literal phrases. In house dataset has been evaluated using Inter rater reliability. The degree to which two or more raters agree is referred to as inter-rater reliability. It addresses the issue of rating system implementation uniformity. We have achieved 82.4% of agreement for training dataset and 85.2% of agreement for testing dataset.

Few examples of Idiomatic sentences for In house dataset is given below:

1. The coach said I have to pull my socks up or I'll lose my spot on the team.
2. The president's illness is preying on everybody's mind.
3. I'm up to my neck in emails and I don't think I can get away at the moment.
4. If all else fails, I can get there by train. It's never fully booked.
5. Kerry has been training hard so she should be firing on all cylinders in the Olympic Games.

We use the grid search method to find weights (Bergstra & Bengio, 2012). Grid search discovers a model's hyper parameters, resulting in the most accurate predictions. The ensemble model's predictions are calculated using the weighted average technique, which is slightly different from the simple averaging method. The latter combine's predictions from each model and take the average, often performing better, overall, than a single model. The weighted average or weighted sum ensemble is a machine learning strategy that aggregates predictions from several models, with each model's contribution weighted according to its competence or performance. Fig 4 shows the pseudo code for the weighted average in the ensemble method, wherein x denotes the weight for the BERT model and Y the weight for the RoBERTa model. POB and POR refer to the predictions of BERT and RoBERTa, respectively, with the values of their predictions ranging from 0 to 1. Using, the weighted average for the predictions is obtained and stored in the variable, final_prediction. Using the argmax function, the respective class labels are identified and stored in the variable output.Where, POB refers to the Predictions of BERT and POR refers to the Predictions of RoBERTa

In next section we outline the experimental setup includes training the baseline models, Bert and RoBERTa, and comparative analysis of the proposed ensemble model with these two baseline models for the task of idiom and literal classification.

4.1. Experimental results

BERT and RoBERTa models were trained by Trofi dataset. Our baseline BERT and RoBERTa models were fine- tuned in 5 epochs, with a maximum sequence length of 120 and a batch size of 16, to predict each class label separately. The models were fine-tuned with

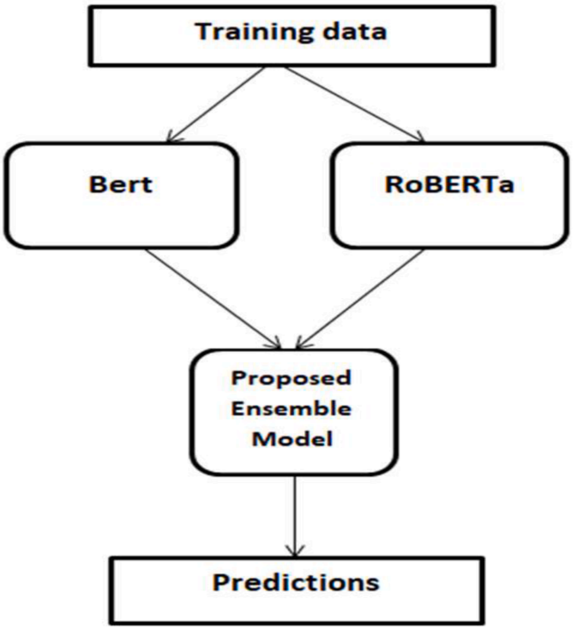
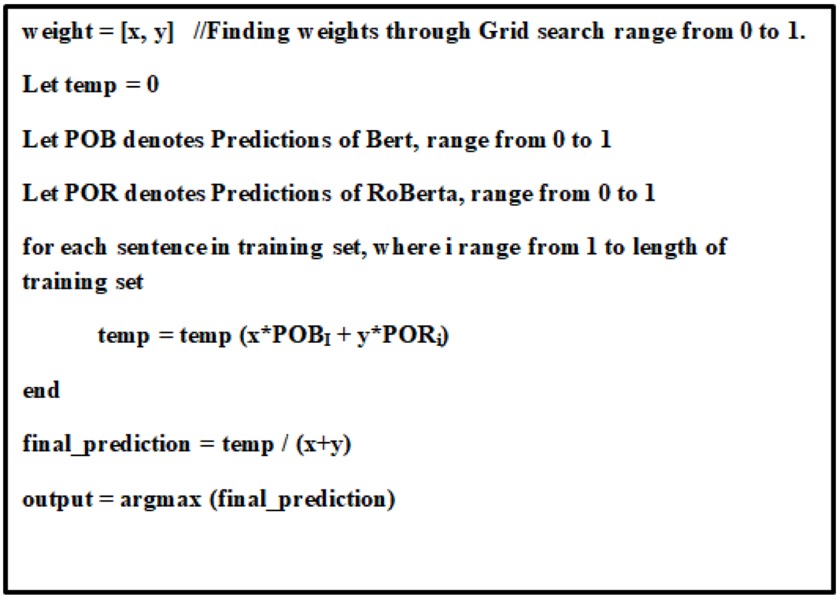


Fig. 3. Proposed Ensemble model Architecture.



```
weight = [x, y]   //Finding weights through Grid search range from 0 to 1.
Let temp = 0
Let POB denotes Predictions of Bert, range from 0 to 1
Let POR denotes Predictions of RoBerta, range from 0 to 1
for each sentence in training set, where i range from 1 to length of training set
        temp = temp ($x*POB_I + y*POR_i$)
end
final_prediction = temp / (x+y)
output = argmax (final_prediction)
```

Fig. 4. Pseudo code for weighted average in ensemble method.

a learning rate of 2 × 10^{-5} and evaluated using the F-score and accuracy metrics. Figs. 5 and 6 depict the confusion matrix of the BERT and RoBERTa baseline models.

F-score and accuracy were calculated using the four true positive (TP), true negative (TN), false positive (FP) and false negative (FN) factors. show precision, recall, F-score and accuracy.

The number of positive class predictions that actually belong to the positive class is measured by precision. The number of positive class predictions made from all positive examples in a dataset is measured by recall. F-score is the harmonic mean of Precision and Recall. It's calculated as the ratio of correct forecasts to total predictions (Banthiya, Kaushiq, Khapre & Shankar, 2021; Egghe, 2008; Hakak et al., 2021; Muthu et al., 2020).Where, TP –No. of idioms correctly classified as idioms

TN – No. of literals correctly classified as Literals

FP – No of literals incorrectly classified as idioms

FN – No of idioms wrongly classified as literals

The confusion matrix for the factors shown in precision and recall is depicted in Figs. 5 and 6.

Fig 5 shows the confusion matrix for Bert. Using, an F-score and accuracy of 0.84 and 0.85, respectively, are achieved.

Fig 6 shows the confusion matrix for RoBERTa. By using F-score achieved is 0.87 and Accuracy achieved is 0.88.

Table 1 shows the evaluation metrics for the models, and Fig. 7 the bar chart visualization for all three models.

Performance and robustness are the two primary reasons for using an ensemble over a baseline model. The proposed ensemble model is the reason for the reduced variance, increased accuracy and F-score. This is because the baseline models, from which the ensemble model was built, were fine-tuned with the TroFi dataset and thus show robust results when tested with our own in house

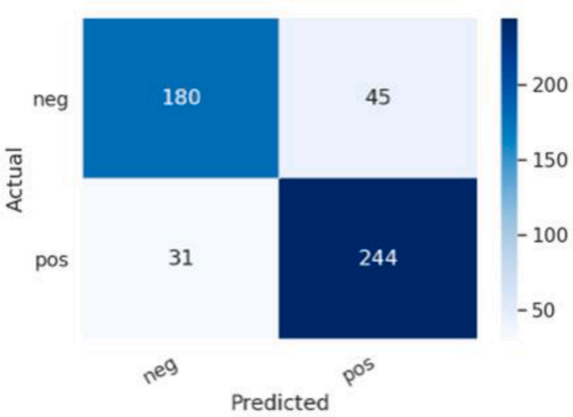


Fig. 5. Confusion matrix for BERT.

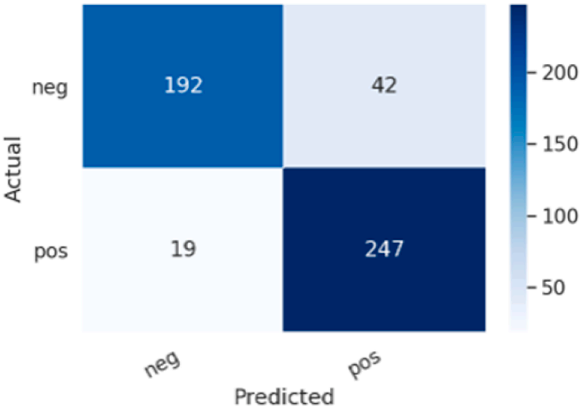


Fig. 6. Confusion matrix for RoBERTa.

Table 1
Evaluation metrics for the models.

Models	F-Score	Accuracy
Bert -base	0.84	0.85
RoBERTa	0.87	0.88
Ensemble Model	**0.89**	**0.90**

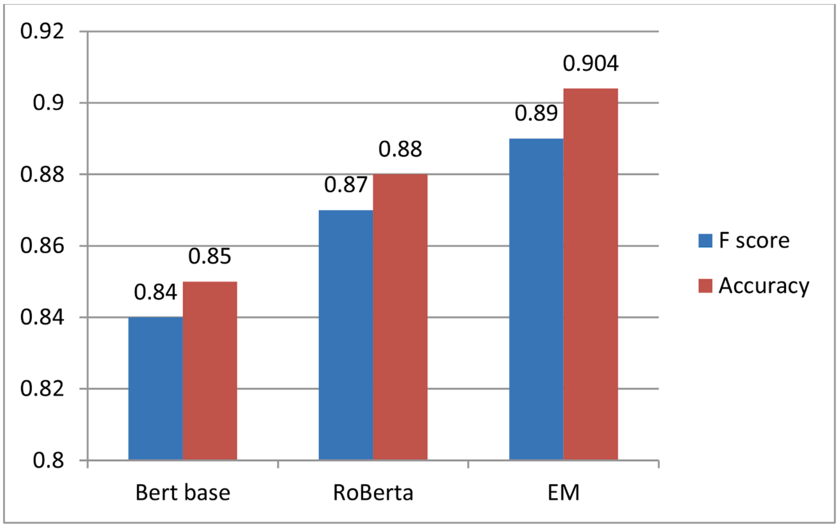


Fig. 7. Bar chart for the metrics.

dataset of 1470 sentences.

Table 1 shows that our ensemble-based model produces better results than the baseline BERT and RoBERTa models. The ensemble model offers a higher 90% accuracy and 89% F-score, compared to those of the baseline BERT and RoBERTa individual models, with a 2% increase in accuracy.

5. Conclusion and future works

This paper has presented an ensemble predictive model for the classification of idiomatic and literal sentences combining two baseline models, BERT and RoBERTa. It achieved a 2% increase in F-score and accuracy, when compared to the baseline models. The proposed ensemble model was fine-tuned with the TroFi dataset and tested with our own dataset, rendering it generic and capable of performing well for any idiom-literal expression dataset, apart from those used here. Proposed model can be used for many NLP applications like Machine Translation, Chatbots etc. Further, and the proposed ensemble model can be used to classify any text, though it is primarily intended for classifying idioms and literals.

It can, therefore, be used for text classification tasks such as sentimental analysis and movie reviews, and enhanced to act as a multi-classifier for text classification tasks demanding multiclass identification. Thus, robust applications can be built atop this classifier, given that the baseline models, BERT and RoBERTa, have been built by Google and Facebook, respectively, with multidimensional texts.

CRediT authorship contribution statement

J Briskilal: Conceptualization, Writing – original draft, Data curation, Visualization, Resources, Validation, Software. **C.N. Subalalitha:** Methodology, Writing – review & editing, Supervision, Project administration, Validation, Investigation.

Declaration of Competing Interest

None.

References

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398.

Annamoradnejad, I., & Zoghi, G. (2020). Colbert: Using bert sentence embedding for humor detection. arXiv preprint arXiv:2004.12765.

Babu, Y.P., & Eswari, R. (2020). CIA_NITT at WNUT-2020 Task 2: Classification of COVID-19 Tweets Using Pre-trained Language Models. arXiv preprint arXiv:2009.05782.

Banthiya, A., Kaushiq, K., Khapre, S., & Shankar, A. (2021). Short text mining method based on sub-semantic space. *Data Driven Approach Towards Disruptive Technologies Studies in Autonomic, Data-driven and Industrial Computing*, 309–323. https://doi.org/10.1007/978-981-15-9873-9_25.

Baruah, A., Das, K., Barbhuiya, F., & Dey, K. (2020). Aggression identification in english, hindi and bangla text using bert, roberta and svm. In *Proceedings of the second workshop on trolling, aggression and cyberbullying* (pp. 76–82).

Bergstra, James, & Bengio, Yoshua Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*.

Birke, J., & Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Blaschke, V., Korniyenko, M., & Tureski, S. (2020)."CyberWallE at SemEval-2020 Task 11: An analysis of feature engineering for ensemble models for propaganda detection." arXiv preprint arXiv:2008.09859.

Chen, X., Leong, C. W., Flor, M., & Klebanov, B. B. (2020). Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing* (pp. 235–243).

Dadu, T., Pant, K., & Mamidi, R. (2020). Bert-based ensembles for modeling disclosure and support in conversational social media text. arXiv preprint arXiv:2006.01222.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Diab, M., & Bhutada, P. (2009). Verb noun construction MWE token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)* (pp. 17–22).

Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing & Management, 44*(2), 856–876.

González-Carvajal, S., & Garrido-Merchán, E.C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems, 117*, 47–58. https://doi.org/10.1016/j.future.2020.11.022.

Jochim, Charles C. (2018). SLIDE-a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Liao, W., Zeng, B., Yin, X., & Wei, P. (2021). An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. *Applied Intelligence, 51*(6), 3522–3533.

Liu, C., & Hwa, R. (2017). Representations of context in recognizing the figurative and literal usages of idioms. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Liu, Y., Liu, B., Shan, L., & Wang, X. (2018). Modelling context with neural networks for recommending idioms in essay writing. *Neurocomputing, 275*, 2287–2293.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. et al. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Mao, J., & Liu, W. (2019). Factuality classification using the pre-trained language representation model BERT. In IberLEF@ SEPLN (pp. 126-131).

Munikar, M., Shakya, S., & Shrestha, A. (2019,. November). Fine-grained sentiment classification using bert. In 2019 Artificial Intelligence for Transforming Business and Society (AITB) (Vol. 1, pp. 1–5). IEEE.

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2020). Detection and Classification of mental illnesses on social media using RoBERTa. arXiv preprint arXiv:2011.11226.

Muthu, B., Cb, S., Kumar, P. M., Kadry, S. N., Hsu, C., Sanjuan, O., et al. (2020). A framework for extractive text summarization based on deep learning modified neural network classifier. *ACM Transactions on Asian and Low-Resource Language Information Processing.* https://doi.org/10.1145/3392048.

Opitz, DavidD., & Maclin, Richard R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 1*, 169–198.

Peng, J., Feldman, A., & Vylomova, E. (2018). Classifying idiomatic and literal expressions using topic models and intensity of emotions. arXiv preprint arXiv:1802.09961.

Risch, J., & Krestel, R. (2020). Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 55–61).

Shaikh, N. (2020). Determination of idiomatic sentences in paragraphs using statement classification and generalization of grammar rules. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation* (pp. 45–50).

Škvorc, T., Gantar, P., & Robnik-Šikonja, M. (2020). MICE: Mining idioms with contextual embeddings. arXiv preprint arXiv:2008.05759.

Spasić, I., Williams, L., & Buerki, A. (2017). Idiom-based features in sentiment analysis: Cutting the Gordian knot. *IEEE Transactions on Affective Computing, 11*(2), 189–199.

Su, T., Macdonald, C., & Ounis, I. (2019). Ensembles of recurrent networks for classifying the relationship of fake news titles. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 893–896).

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?. In *China National Conference on Chinese Computational Linguistics* (pp. 194–206). Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. et al. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998–6008).

Verma, R., & Vuppuluri, V. (2015). A new approach for idiom identification using meanings and the Web. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 681–687).