An Ensemble Method to Classify Telugu Idiomatic Sentences using Deep Learning Models

J Briskilal

Department of Computing Technologies
SRM Institute of Science and Technology

Kattankulathur, Chennai – 603203.

briskilj@srmist.edu.in

Ch V M Sai Praneeth

Department of Computing Technologies
SRM Institute of Science and Technology

Kattankulathur, Chennai – 603203.

cs9184@srmist.edu.in

Ch Chaitanya

Department of Computing Technologies
SRM Institute of Science and Technology

Kattankulathur, Chennai – 603203.

cc4789@srmist.edu.in

M Jaya Karthik

Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Chennai – 603203.

mk3174@srmist.edu.in

P Purnachandra Reddy

SRM Institute of Science and Technology

Kattankulathur, Chennai – 603203.

pr4494@srmist.edu.in

2023 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-9849-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICT57646.2023.10134038

***Abstract*— Text classification is a requirement for every text processing application because the web contains a vast amount of text data. Intent detection, information extraction, sentiment analysis, and spam detection involves text categorization. Since text classification uses idioms, metaphors, and polysemic words, intent detection can be difficult. It is challenging to automatically identify idioms in Natural Language Processing applications such as Information Retrieval, Machine Translation, and chatbots. In all these applications, automatic idiom recognition is crucial. In this work, idiomatic and literals sentences are being classified. Idioms are typical expressions with new meanings. This research proposes an ensemble model using pre-trained deep learning models to make model with more predictive nature. The models are trained and tested using in-house dataset. Moreover, an in-house dataset that contains 1040 idiomatic and literal sentences is suggested. The experimental results demonstrate the effectiveness of the proposed approach, achieving an accuracy of 86% on the test dataset.**

***Keywords*— Natural language processing, Cross-lingual Language Model (XLM)-Roberta, Multilingual-Bert, Idiom and literal classification, Text classification, deep learning and ensembled models.**

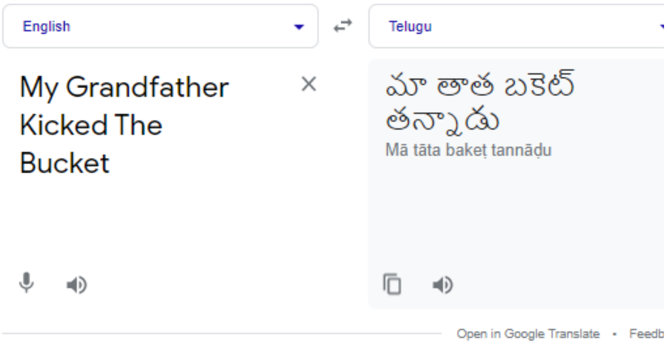
I. INTRODUCTION

Text classification is a term that can be used to describe both the labelling and grouping of text data as well as the classification process itself. The process of text classification involves organizing or labelling textual data. One additional way to think about this is as the process of grouping together the various kinds of textual data. The natural language processing (NLP) must invariably incorporate this aspect as an essential component. As a result of living in the digital age, users are frequently exposed to text in several formats. This can be both beneficial and detrimental to the learning. This type of content includes, but is not limited to, the text that appears on the social media accounts, in adverts, websites, digital books, and other types of digital media. Classification is a tool that can be of great assistance in determining the content of this text data because most of it is unstructured. The classification of texts can be used in a wide variety of situations. The detection of spam, the conduct of sentiment analysis, the labelling of topics, the identification of languages, the categorization of online content, and the determination of the author's intent are some of the applications of this technology. Labeling topics, identifying languages, labelling topics, and classifying online content are some examples of additional applications. In this piece of research, the process of text classification is accomplished with the assistance of literals and idioms. Idiomatic and literal classifications of words and phrases are frequently used in natural language processing applications to organize and categorize words and phrases, such as Machine Translation and Information Retrieval (IR) systems.

The inherent complexities that are present in language interpretations continue to be the focus of a significant amount of research, and natural language is still the subject of this research. This situation becomes even more precarious when computers are used to perform language interpretations automatically. This calls for performing an automatic disambiguation of the text to derive the meaning that was intended. To make one such attempt, the primary focus of this work is on classifying idioms in accordance with the literal expressions that they are supposed to correspond to. It has been determined that the task at hand is one of text classification, and pre-trained deep learning models and ensembled models have been utilized to carry out the task to carry out the implementation.

A phrase is literal if it contains words that are the same as those found in an idiomatic phrase; however, the literal phrase communicates a meaning that is derived from the words themselves rather than an idiomatic meaning. Words that are used in an idiomatic phrase may also be included in a literal phrase; Yet rather of conveying the terms' figurative meaning, they are used in their literal sense. A phrase is idiomatic if it contains words that do not have their literal meaning preserved.

Challenges encountered due to the lack of automatic idiom recognition:

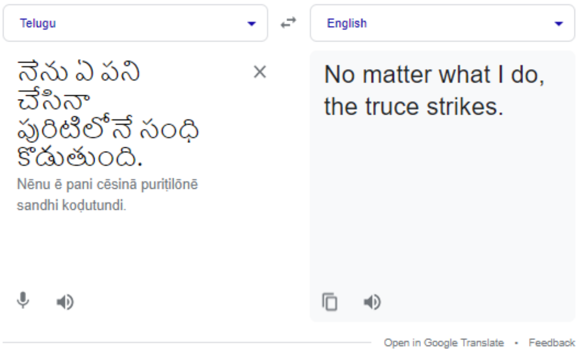


My Grandfather Kicked The Bucket

మా తాత బకెట్ తన్నాడు

Mā tāta baket tannāḍu

Fig.1: Example 1 of mistranslation due to the lack of automatic idiom recognition



నేను ఏ పని చేసినా పురిటిలోనే సంధి కొడుతుంది.

Nēnu ē pani cēsinā puriṭilōnē sandhi koḍutundi.

No matter what I do, the truce strikes.

Fig.2: Example 2 of mistranslation due to the lack of automatic idiom recognition

How a machine translation system can mistranslate has been illustrated in figure 1. "My Grandpa Died" must appear as the proper translation provided by the machine translator. The translation is done directly based on the words in the sentence because the translator is unable to determine if the sentence is idiomatic or not. Like example 1- The translator has given the incorrect translation in figure 2 - Which should be "Whenever I start working on something, it leads to failure." Due to a lack of automatic idiom identification, this is one of the areas where these applications do not perform as intended. These are the challenges for the creation of automatic idiom recognition.

Example idiomatic sentences are given below:

1. నేను ఏ పని చేసినా పురిటిలోనే సంధి కొడుతుంది.

   Nenu ey pani chesina puriṭilone sandhi koḍutundi

   Whenever I start to work on something it leads to failure.

2. రవి గాడు ప్రస్తుతం వదిలేసి పగటి కలలు కంటున్నాడు.

   Ravi gaḍu prastutam vadilesi pagaṭi kalalu kanṭunnaḍu

   Leaving the present, ravi is dreaming about future.

3. రాము ఆలస్యంగా రావడంతో పంతులు చేతిలో అక్షింతలు పడ్డాయి.

   Ramu alasyanga ravadamtho panthulu chetilo akshintalu paddayi

   Ramu was scolded by his teacher because of his late presence.

4. వాడు పక్కదారి పట్టించడంలో దిట్ట.

   Vadu pakkadari pattinchadamlo ditta

   He is very talented in making fool of others.

II. LITERATURE SURVEY

The literature review was carried out concurrently along two separate dimensions at the same time. The first study approaches the issue from the standpoint of idiom recognition, and the second study approaches it from the standpoint of utilizing machine learning techniques to classify texts according to whether they include idioms. Due to these two elements, a deeper understanding of the traits and classifications of machine learning algorithms that are applied to text categorization and idiom recognition has been gained. These algorithms are used to recognize phrases and text. Idiom recognition is made possible by these algorithms, which can be found in most modern computers.

A. Text classification works using Roberta model

RoBERTa and SVM classifier usage in the context of aggressiveness detection-based classification [8] was discussed by Arup et al. in 2020. A dataset that can be used to identify aggressive conduct has different interpretations across all three languages—English, Hindi, and Bangla. The following interpretations are listed. The data in consideration can be compared to one another. For the English subtasks, the RoBERTa model outperformed the SVM classifier with an F-score of 80%. This result was achieved through the application of the RoBERTa model [1]. (Baruah, Das, Barbhuiya & Dey, 2020). During their investigation into the aspect category sentiment analysis, Wexiong et al. (2020) made extensive use of the RoBERTa model as the primary instrument that served as their guide. The classification was completed with the assistance of a dataset from AI Challenger that was accessible to the public and included fine-grained sentiment analysis [9] (Liao, Zeng, Yin & Wei, 2021). For classification, the publicly available dataset (fine-grained sentiment analysis) that was supplied by AI challenger was utilised. To acquire the outcomes that were aimed for, this action was taken. Because they used Roberta [2], Ankit et al. (2020) was successful in locating and classifying posts on social media that were related to mental illness. The dataset that was used to determine the five distinct class labels was built with a total of three thousand unique posts that were culled from a variety of subreddits. These posts were used to build the dataset. The dataset that was compiled utilized these posts in various ways. When compared to the BERT and LSTM [16] models, RoBERTa demonstrated performance that was clearly more effective than either of those models.

B. Works on classification of idioms

To identify and classify idiomatic expressions and literal expressions into the categories that are most suited for them based on the meaning of the expressions, Peng [3] suggested a method. A bag-of-words subject representation was used to categorize phrases as literal expressions or idioms depending on the topics that were considered, and this was done to determine which category they most effectively fit into. This was done to classify sentences as either literal expressions or idioms, and this was done to determine which category they most effectively fit into. It was necessary to do this to categorize sentences as either literal expressions or idioms [4], and it was also necessary to do this to determine which category sentences most naturally fit into. This was done so that sentences could be classified as literal expressions, idioms, or some other type of expression.

This method of classification made use of four separate datasets, including BlowWhistle (which contained a total of seventy-eight sentences, including 51 literals and 27 idioms), MakeScene (which contained a total of fifty sentences, including 20 literals and 30 idioms), LoseHead [5](which contained a total of forty sentences, including 19 literals and 21 idioms), and TakeHeart (which contained a total of eighty-one sentences, including 61 idioms. By employing idiom-based features as the data source, Irena and her colleagues (2017) proposed an automated way to enhance sentiment analysis [15]. Irena came up with this strategy. For classification, a dataset that is universally acknowledged as serving as a standard for the industry was utilized. In addition to this, the analysis consisted of a set of rules to recognize idioms and sentiment polarities, both of which were taken into consideration (Spasic, Williams, and Buerki, 2017).

When classifying idioms and literals that appear within paragraphs, Naziya (2020) utilized both a rule-based generalization as well as a context-based classification. This was done to categorize literals and idioms that are found within paragraphs [10]. (Shaikh, 2020) In the research that Liu [11] and his colleagues carried out and published in 2018, it was suggested that an approach that is based on neural networks be used to make recommendations regarding the utilization of idioms in the writing of essays. Specifically, it was suggested that idiomatic expressions be used more frequently in academic writing. Idioms were specifically mentioned as something that should be utilized more frequently in academic writing. Before the results were presented, the level of similarity that existed between the given context and candidate idiom [10] recommendations were analyzed and computed (Liu, Liu, Shan & Wang, 2018). A token-level metaphor classification procedure was carried out by Xianyang et al. (2020) utilizing a BERT model on the TOEFL corpus as well as the VUA (VU Amsterdam Metaphor corpus). Xianyang et al. oversaw this process. This technique was applied to these two corpora separately (Chen, Leong, Flor & Klebanov, 2020).

A supervised ensemble model that combines late and early fusion strategies to classify idiomatic expressions in addition to literal expressions was proposed by Liu et al. (2017). This model was used to classify expressions into the proper groups (Liu & Hwa, 2017). Crowdsourcing may be a useful technique for gathering the emotional annotations of idiomatic idioms, according to the research conducted by Charles and his colleagues. Charles did this investigation (2018). This procedure resulted in the creation of the Emotion Lexicon of Idiomatic Expressions [8] (SLIDE), which is notably more thorough than previous lexicons (Jochim et al., 2018). Mona et al. came up with the idea of using supervised learning as an approach to the classification of multiword expressions as opposed to literals (2009). This investigation made use of the VNC-Tokens dataset, which is a construction involving a verb and a noun. The authors were successful in accomplishing this goal thanks to the utilization of the dataset.

C. Use of ensemble models for text classification

A bagging-based ensemble model [15] was suggested by Julian et al. (2020) to classify 6000 hostile and offensive social media postings using several upgraded BERT models

(Risch & Krestel, 2020). According to Tadej et al. (2020), an ensemble BERT and ELMo model should be used to categorize idioms and literal utterances. Contextual embedding should underpin this method. The monolingual classification was completed using a dataset in Slovenia (Skvorc, Gantar, & Robnik-Sikonja,2020). Using ensemble models, Harita Reddy, Namratha Raj, Manali Gala, and Annappa Basava (2020) created text categorization [19] to identify false news. Automated Text Tagging of Arabic News Items [6] Using Ensemble Deep Learning Models was created by Ashraf Elnagar, Omar Einea, and Ridhwan Al-Debsi (2019). Using knowledge-enabled BERT in deep learning, S. Abarna, J. I. Sheeba, and SP. Devaneyan created an ensemble model for the categorization of idioms and literal text [7] in 2022. Using BERT and RoBERTa, J. Briskilal and CN. Subalalitha created an ensemble model [17] for the classification of idioms and literal texts in Tamil.

Idioms and literal texts in Telugu were not categorised utilising ensemble techniques, according to the available studies. Ensemble approaches combine numerous models rather than using just one to increase the accuracy of outcomes in models. The combined models considerably improve the accuracy of the outcomes. Regression and classification are perfect applications for ensemble methods because they increase model accuracy by lowering bias and variance. This study categorises literals and idioms using an ensemble model using both deep learning and machine learning models that have been previously trained.

III. Proposed Work

Automatic categorization of news articles, sentiment text classification, predicting movie reviews, hate speech detection on Facebook, and many other tasks have been completed in the Telugu language; However, Idiomatic Sentence Classifier has not been developed yet. The text classifier is used as a means of determining whether a Telugu text contains literal translations or idiomatic translations. Because Telugu is a language with few available resources, own in-house dataset has been compiled. Through the utilization of the kappa score and the IRR, the quality of the dataset is enhanced. The ensemble algorithm is implemented in the proposed system to achieve a higher level of precision than the current system.

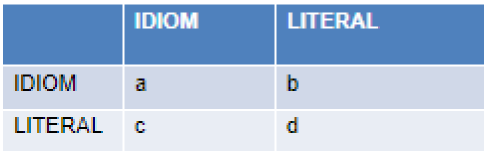
XLM-R, which stands for XLM-roBERTa, which stands for Unsupervised Cross-lingual Representation Learning at Scale, is a cross-lingual sentence encoder that is scaled. It was trained using 2.5 terabytes of data spanning hundred different languages that was filtered from data obtained from Common Crawl. The results that XLM-R achieves on a variety of cross-lingual benchmarks are at the cutting edge of current research. Even while earlier research in this area has shown that multilingual masked language models can improve cross-lingual comprehension, models like XLM and multilingual BERT are limited in their capacity to recognize useful representations for languages with limited resources. The XLM-R is an advancement over preceding techniques in several ways. In comparison to the previous state of the art, which was trained in 15 languages, the performance of XLM-R on low-resource languages improved by 2.3 percent and 5 percent over XNLI on Swahili and Urdu, respectively.

IV. Implementation

The pre-trained deep learning models and ensemble models were used in this work to classify the idioms and literals that were contained inside the text that was provided as input. A compilation of 1040 sentences that encompasses both the literal and figurative applications of several idiomatic idioms. Below figures depicts the suggested experimental set-up for idiomatic expression and literal classification. The features are then given to the classifiers once the input text has been preprocessed and the features necessary for classifying the text have been extracted.

A. In-House Dataset Creation

Since Telugu language is a low resource language. Own in-house dataset has been created, in which idiomatic sentences are created using idioms in Telugu language and literal sentences taken from websites like Wikipedia which are available in Telugu. The in-house dataset includes 1040 idiom and literal sentence examples. There are a total of 520 idiomatic sentences and the same number of literal sentences. Both the text sentences and the annotations for this research were authored and annotated by domain experts. Since own in-house dataset is created, the quality of the dataset must be determined. For that, a Google form is developed to collect the annotations from other people depending on their perspectives. To determine how accurate the dataset is, the kappa score is used in conjunction with the inter-rater reliability by considering annotations given by us and others. Kappa's score was found using the confusion matrix. The below figure depicts the confusion matrix.



	IDIOM	LITERAL
IDIOM	a	b
LITERAL	c	d

Fig.3: confusion matrix

$$Kappa\ score = \frac{OA - CA}{1 - CA}$$

The kappa score is found using OA and CA where OA is observed agreement and CA is chance agreement and the formula is shown above. Here Observed agreement is equals to (a+d)/(a+b+c+d), while chance agreement is equals to ((a+b)*(a+c)) + ((b+d)) *(d+c)/(a+b+c+d))$^{2.}$ Inter-rater reliability was found to be 92%, with a kappa score of 0.8394 and an observed agreement of 0.92. The chance of agreement was found to be 0.5016. As a result, the level of agreement for the dataset is extremely high or even perfect.

B. Selection of models

Since the machine only understands numbers, the tokenizer is used to translate the given Telugu sentence into numbers. While selecting models from machine learning, the Bag of Words tokenizer is used to tokenize the Telugu sentences, after that the top machine learning algorithms available are trained, and the models giving accuracy greater than 75% have been considered. After putting the models to the test, three algorithms—SVM, Logistic Regression, and Naïve Bayes that offered good accuracy are considered. Additionally, these models are combined using ensemble methods in order to produce a better model with low bias and variance. Dravidian languages are among the more than one hundred languages that the deep learning models XLM-roBERTa and m-BERT have been trained on. Thus, that these models may be trained more effectively than machine learning models and quickly understand Telugu. For this reason, these models are used to create ensemble models using ensemble techniques.

C. Training and testing the models

8:2 split ratio is used to divide the entire dataset into two parts: a training dataset and a testing dataset. XLM-roBERTa, m-BERT, and an ensemble model consisting of machine learning algorithms (SVM, Logistic Regression, Naïve Bayes) and simple average ensemble model using XLM-roBERTa and m-BERT are then trained. This allowed to compare the performance of each model on the test data. The models are assessed using the test dataset by determining its accuracy, precision, f1 score, and recall through the process of comparing the predicted outputs and supplied outputs with the assistance of a confusion matrix. This evaluation is carried out with the help of the test dataset. The models block diagram can be seen in the figures below.

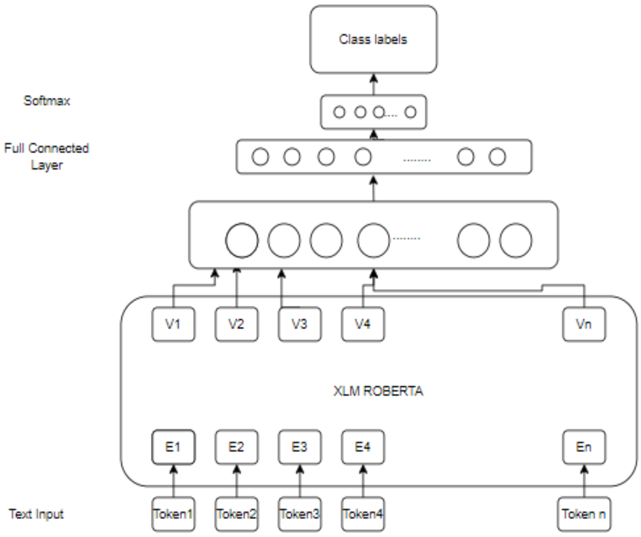


Fig.4: Architecture of XLM-roBERTa model

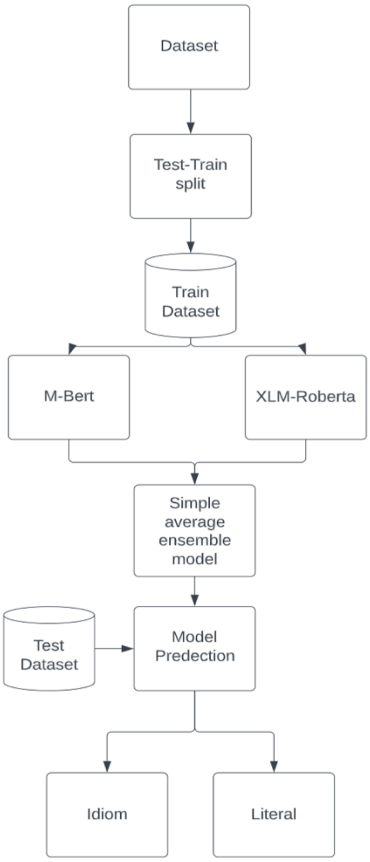


Fig.5: Architecture of ensembled model for deep learning models

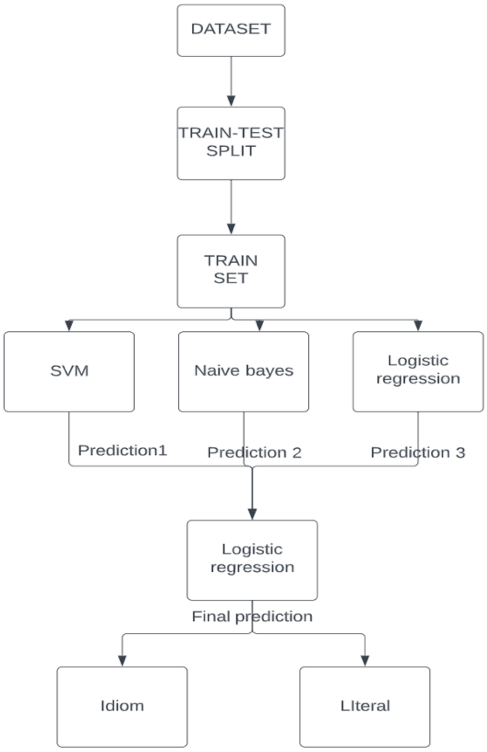


Fig.6: Architecture of stacked ensembled model for ML models

The machine learning algorithms are trained over the dataset after using bag of words to vectorize the dataset so that the machine learning algorithms understand the Telugu dataset. After testing the machine learning algorithms, the top three algorithms are used to create a stacked ensemble model and trained it over the dataset. The sentence piece tokenizer, Adamw optimizer, cross entropy loss function, batch size 32 and 3 epochs are used when training XLM-roBERTa and m-BERT. The simple average ensemble model is made by combining the individual results from both XLM-roBERTa and m-BERT. The observations made regarding the outcomes of the experiment are covered in the next section.

V. Results discussion

The model was validated using an in-house dataset of 1040 sentences, 520 of which were idiomatic sentences and the remaining 520 were literal sentences. Performance of the model was measured using four criteria: precision, f1-score, recall, and accuracy.

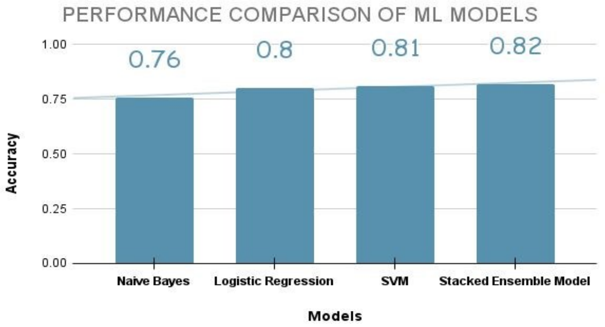


Fig.7: Accuracy comparison between ML and its stacked ensemble model

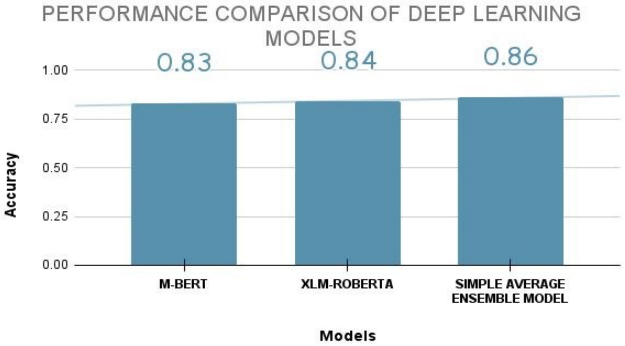


Fig.8: Accuracy comparison between DL and its stacked ensemble model

Table 1: Comparison between DL and Ensemble models

MODEL	PRECISION	RECALL	F1-SCORE	ACCURACY
Stacked ensemble using ML	0.86	0.8	0.8	0.82
XLM-roBERTa	0.87	0.85	0.84	0.84
m-BERT	0.86	0.84	0.83	0.83
Simple average ensemble using DL	0.88	0.87	0.86	0.86

Table 1 shows that the simple average ensemble model outperformed pre-trained deep learning algorithms and stacked ensemble using ML in terms of precision, recall, f1-score, and accuracy. According to Fig 6, the stacked ensemble model outperforms well compared to Naïve Bayes, Logistic regression and SVM with accuracies of 0.82,0.76,0.8 and 0.81 respectively. From Figure 7 the simple average ensemble model generates superior results than XLM-roBERTa and m-BERT. Based on results the stacked ensemble model is predicting nearer to m-BERT. Because of the 160 GB of text used to pre-train the basic Roberta model, 16 GB of which came from the Books Corpus and the English Wikipedia that is utilized in BERT, The extra data included the Common Crawl News dataset, which has 63 million items and 76 gigabytes, the Web text corpus, and the Common Crawl Stories dataset (31 GB) makes the simple average ensemble model performed well. In addition to this, the XLM-roBERTa model makes use of dynamic masking, and the m-BERT model makes use of static masking. Both of these contribute to the robustness of the simple average ensemble model.

VI. Conclusion

The ensemble prediction model for the categorization of literal and idiomatic phrases proposed in this research combines the m-BERT and XLM-roBERTa baseline models. When compared to the baseline models, the accuracy has increased by 2%. Many NLP applications, including Machine Translation and chatbots can make use of the proposed approach. This model can be included in machine translation so that it can classify a statement as literal or idiomatic and translate it accordingly. By being included in the NLP applications, this model will raise the value of those applications. To make this model more predictive, the dataset size must be increased, so that it can be trained over more varieties of idiomatic and literal sentences.

VII. References

[1] Feldman A, Peng J (2013) Automatic detection of idiomatic clauses. In: International conference on intelligent text processing and computational linguistics. Springer, Berlin, Heidelberg.

[2] Fazly A, Cook P, Stevenson S (2009) Unsupervised type and token identification of idiomatic expressions. Comput Linguist 35(1):61–103.

[3] Peng J, Feldman A, Vylomova E (2018) Classifying idiomatic and literal expressions using topic models and intensity of emotions.

[4] Singh G et al (2019) Comparison between multinomial and Bernoulli Naïve Bayes for text classification. In 2019 International conference on automation, computational and technology management (ICACTM). IEEE

[5] Bahassine S et al (2020) Feature selection using an improved Chi-square for Arabic text classification. J King Saud Univ-Comput Inf Sci 32(2):225–231

[6] Elnagar, Ashraf, Ridhwan Al-Debsi, and Omar Einea. "Arabic text classification using deep learning models." *Information Processing & Management* 57.1 (2020): 102121.

[7] Abarna, S., J. I. Sheeba, and S. Pradeep Devaneyan. "An ensemble model for idioms and literal text classification using

knowledge-enabled BERT in deep learning." *Measurement: Sensors* 24 (2022): 100434.

[8] Baruah, Arup, et al. "Context-aware sarcasm detection using bert." *Proceedings of the Second Workshop on Figurative Language Processing*. 2020. Rother, K., Rettberg, A.: Ulmfit at germeval-2018: A Deep Neural Language Model for the Classification of Hate Speech in German Tweets. pp. 113–119 (2018)

[9] Liao, Wenxiong, et al. "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa." *Applied Intelligence* 51 (2021): 3522-3533.

[10] Shaikh, Naziya. "Determination of idiomatic sentences in paragraphs using statement classification and generalization of grammar rules." *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*. 2020.

[11] Adhikari, Ashutosh, Achyudh Ram, Raphael Tang, and Jimmy Lin. "Docbert: Bert for document classification." *arXiv preprint arXiv:1904.08398* (2019).

[12] Xiaoyu Luo, June 2021 Efficient English text classification using selected Machine Learning Techniques.

[13] Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi Sentiment analysis and classification of Indian farmers' protest using twitter data (2019).

[14] Ningfeng Sun and Chengye Du News Text Classification Method and Simulation Based on the Hybrid Deep Learning Model (2021).

[15] Briskilal, J., & Subalalitha, C. N. (2022). An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. Information Processing & Management, 59(1), 102756.

[16] Briskilal, J., & Subalalitha, C. N. (2021). Classification of Idioms and Literals Using Support Vector Machine and Naïve Bayes Classifier. In Machine Vision and Augmented Intelligence—Theory and Applications (pp. 515-524). Springer, Singapore.

[17] Briskilal, J., & Subalalitha, C. N. (2022). Classification of Idiomatic Sentences Using AWD-LSTM. In Expert Clouds and Applications (pp. 113-124). Springer, Singapore.

[18] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

[19] Reddy, H., Raj, N., Gala, M., & Basava, A. (2020). Text-mining-based fake news detection using ensemble methods. *International Journal of Automation and Computing*, *17*(2), 210-221.