T20NGD: Annotated corpus for news headlines classification in low resource language,Telugu.

Chindukuri Mallikarjuna
NIT-Tiruchirappalli,Tamilnadu,India
malli.chindukuri@gmail.com

Sangeetha Sivanesan
NIT-Tiruchirappalli,Tamilnadu,India
sangeetha@nitt.edu

Abstract

News classification allows analysts and researchers to study trends over time. Based on classification, news platforms can provide readers with related articles. Many digital news platforms and apps use classification to offer personalized content for their users. While there are numerous resources accessible for news classification in various Indian languages, there is still a lack of extensive benchmark dataset specifically for the Telugu language. Our paper presents and describes the Telugu20news group dataset, where news has been collected from various online Telugu news channels. We describe in detail the accumulation and annotation of the proposed news headlines dataset. In addition, we conducted extensive experiments on our proposed news headlines dataset in order to deliver solid baselines for future work.

1 Introduction

In the digital era of today, the quantity of news content is overwhelming. Globally, thousands of articles are published every minute on a wide variety of topics, ranging from international politics to local athletics. As the quantity of news grows exponentially, the need for effective and efficient classification of news becomes paramount. This ensures that readers are presented with content pertinent to their interests and that platforms can manage and distribute their vast content libraries effectively.

News categorization is an important aspect of news research (Katari and Myneni, 2020). We utilize the information provided by the news to categorize them into various groups (Wang et al., 2022). Some work has been done on news classification; for instance, there are news datasets like 20NEWSGroup(Dua and Graff, 2017) and AG News (Zhang et al., 2015). Numerous methodologies have been devised in recent studies to evaluate the classification of news texts and the difficulties associated with it. However, the majority of these works are restricted to a small number of languages, with English being the predominant language. In spite of the fact that Telugu is the fourth most spoken language in India and the sixteenth most spoken language in the globe (Regatte et al., 2020), the classification of news text in Telugu remains unexplored. There have been some resources created in Telugu supported for various NLP tasks such as sentiment analysis(Mukku and Mamidi, 2017; Parupalli et al., 2018; Gangula and Mamidi, 2018), sentiment emotion(Marreddy et al., 2022), hate speech detection (Marreddy et al., 2022), and for code-mixing (Kusampudi et al., 2021). However, there is no gold standard benchmark dataset for supporting Telugu news headlines text classification task. Therefore through this paper, we present a benchmark dataset (T20NGD) for Telugu news headlines text classification that has been meticulously annotated to serve as a gold standard for news headlines classification.

The structure of the paper is arranged as. In part 2, we review prior Telugu text classification approaches, existing Telugu corpus designed for NLP tasks, and recent improvements. In Section 3, we outline our corpus and annotation methodology. Sec. 4 briefly describes the models evaluated over the proposed dataset. Section 5 outlines experiments and results discussion made in section 6. Finally, our conclusions are presented in section 7.

2 Related study

Numerous approaches have been employed to classify news headlines in Indian languages, with a limited number of studies focusing on the Telugu language. This section provides an overview of the

Table 1: Comparision of previous telugu news classification datasets with proposed dataset.

Reference	Source	Size	Approach	Classes
Sudha et.al (Sudha et al., 2021)	Daily Haunt	600	SVM,MNB	5
Murthy et.al (Murthy, 2003)	Eenadu	794	Naive Bayes	4
Veerraju et.al (Gampala et al., 2021)	–	5000	SVM,Naive Bayes,MLP	5
Swapna et.al (Narala et al., 2017)	Wikipedia,news channels	1169	Rule-based N-gram	7
Sravya et.al (Sravya et al., 2022)	Newspapers	17600	CNN,LSTM,BiLSTM	5
This Paper(T20NGD)	Online Telugu news Channels	29744	Pre-trained Models	20

strategies and approaches employed in the context of news headline classification specifically in the Telugu language.

Murthy et.al (Murthy, 2003) presents a study where the Naive Bayes classifier is employed for supervised classification of Telugu news articles across four primary categories including Business, Cinema, Politics, and Sports over the corpus of 796 news Telugu documents. Kanaka durga et.al (Durga and Govardhan, 2011) proposed ontology-based text classification of Telugu documents. Vishnu et.al (Murthy et al., 2013) examines the effectiveness of various classification methods utilizing different term weighting techniques to determine the accurate classifier among Naive Bayes (NB), Support Vector Machine (SVM), and K Nearest Neighbor (KNN) for Telugu text classification. Experiments were carried out on a corpus consisting of 800 news articles categorized into class labels such as politics, science, sports, culture, and health and results proved that SVM outperformed NB and KNN. Narayana et.al (Swamy et al., 2014) evaluated the performance of multiple ML algorithms like NB, KNN, and decision tree against the corpus collected from south Indian languages like Tamil, Kannada, and Telugu of each 100 documents. Swapna et.al (Narala et al., 2017) applied a variant of KNN ML algorithm for the classification of 1500 Telugu documents annotated with news, songs, stories, politics, rivers, sports, and literature class labels. Sunil et.al (Gundapu et al., 2020) designed a multi-channel LSTM-CNN approach for Technical Domain Identification in the Telugu language. Sravya et.al (Sravya et al., 2022) applied two variants of SVM model(SVM-linear and SVM-Polynomial) and deep learning models such as Conv1D, LSTM, GRU, BiLSTM, BiGRU for classification of Telugu news articles headlines. These experiments were conducted with the corpus of five categories of news headlines including sports, entertainment, editorial, business, and nation. Experimental results proved that, GRU with Fasttext word embeddings gain good results as 70.92% accuracy which outperformed other ML and DL models.

Based on our review of the literature, it was observed that all previous research endeavors in the field of Telugu text categorization have focused only on either the document level or the sentence level. These works were executed with a limited corpus. Furthermore, it failed to encompass a wide range of categories. Through this research, we present a dataset containing a large number of instances and covering a broad spectrum of news categories. The contributions made in this work are listed as follows.

- Constructing a Telugu news headlines dataset with a large number of instances collected from various domains.
- Design a methodology that classifies Telugu news headlines with maximum accuracy.
- Finetuning different multi and monolingual pre-trained models for Telugu news headlines classification.
- Analyze the experimental results and identify the specific cases where the models fail to accurately classify the news headlines text.

3 Dataset Construction

The process of constructing the dataset involved distinct phases: the collection of news headlines, pre-processing and the subsequent annotation of these pre-processed headlines with their corresponding labels.

3.1 Data Collection

Text classification starts with collecting data, which is a very important step, especially when looking at the huge world of online news sources.We gathered news headlines from multiple sources of popular online news media channels[1] that disseminate news in the Telugu language. The Telugu news headlines were extracted using Beautiful Soup, a popular Python tool commonly employed for web scraping tasks to retrieve data from HTML and XML documents. To efficiently accomplish our goal, we implemented a keyword-based search methodology to gather a wide range and diversified news headlines from multiple online Telugu news channels.

3.2 Pre-processing

After the completion of the data collection phase, we moved on to the pre-processing phase of dataset construction. Because the raw data contained unwanted characters and phrases. For instance, the sentences are prefixed with English tags, publishing dates, and time values. We have automated the elimination of these words and characters. The major pre-processing steps which were carried out in this phase were briefly described as follows.

- **Removing redundant news headlines:** We removed news headlines from the raw data which were repeated twice.
- **Elimination of non-Telugu words and unnecessary symbols:** We eliminated non-Telugu words that were prefixed and followed after each Telugu news headline. In addition, we also removed unnecessary symbols.
- **Removing shorter news headlines:** We excluded news headlines from the corpus that have a length of fewer than three words.

After completion of the pre-processing phase, we had clean corpus with 29744 Telugu news headlines. Table 2 presents the statistics of the proposed corpus.

Table 2: Statistics of the Corpus

Description	Count
Total Sentences	29744
Total Number of Words	271083
Average Sentence Length	72
Maximum Sentence Length	791
Minimum Sentence Length	11

3.3 Annotation

Once after the completion of the pre-processing phase, we proceed with the annotating process. In this section, we describe the annotation process of the proposed dataset. The annotation process includes multiple things such as annotation schema design, selection of annotators, annotation procedure, validation of labels and finally assessment of annotation quality. We described each step in the annotation process briefly in subsequent sections.

3.3.1 Annotation schema design

The news headlines in our dataset were gathered from many domains. To meet this requirement, we developed a schema with 20 class labels including Agriculture, Business, Crime, Development, Education, Employment, Entertainment, Environment, Fashion, Health, International, Judicial, Literature, Parliamentary, Politics, Science and Technology, Spiritual, Sports, Tourism, and Women Empowerment. The labels are briefly defined as follows:

- **Agriculture:** The news headlines related to agriculture-related activities like forming, cultivation, and suggestions given by the domain experts were labelled with the "Agriculture" class label.
- **Business:**The news headlines refer to the economy, stock market trend, financial results of various corporate companies, Govt fiscal policies labelled with the "Business" class label.
- **Crime:** The news headlines referring to robberies, murders, frauds, and other significant events were annotated with the "Crime" class label.
- **Development:** The news headlines refer to the inauguration of new projects, laying the

[1] https://www.eenadu.net
https://www.andhrajyothy.com
https://www.bbc.com/telugu
https://tv9telugu.com
https://telugu.oneindia.com
https://telugu.news18.com
https://ntvtelugu.com
https://zeenews.india.com/telugu
https://10tv.in/
https://sakshi.com
https://www.v6velugu.com
https://telugu.hindustantimes.com
https://telugu.economictimes.com/

foundation for new projects were labelled with the "Development" class label.

- **Education:** The news headlines refer to the admission notification, scholarships, schemes, education loans, entrance test schedules and results labelled with the "Education" class label.
- **Employment:** The news headlines which refer to job notifications and recruitment results were labelled with the "Employment" class label.
- **Entertainment:** The news headlines refer to television shows, films, music, theatre, celebrities, and other forms of entertainment that were labelled with the "Entertainment" class label.
- **Environment:** The news headlines referring to natural disasters, climate change, pollution,and environmental govt policies were labelled with the "Environment" class label.
- **Fashion:** The news headlines refer to the latest trends, announcements, fashion events, new brand items information, and fashion tips labelled with the "Fashion" class label.
- **Health**: The news headlines referring to health tips, disease outbreaks, healthcare policies, nutrition and health advisories were labelled with the "Health" class.
- **International:** The news headlines refer to the global wide events and summits were labelled with the "International" class label.
- **Judicial**: The news headlines refer to the news, judgements, changes in laws, and judicial policies were labelled with the "Judicial" class.
- **Literature:** The news headlines referring to literature events, awards and meetings were labelled with the "Literature" class label.
- **Parliamentary:** The news headlines referring to debates, discussions, proposed laws, and parliamentary proceedings were labelled with the "Parliamentary" class label.
- **Politics:** The news headlines which cover political party meetings, comments, events, and activities were labelled with the "Politics" class label.
- **Science and Technology:** The news headlines which cover recent advancements, recent discoveries, breakthroughs, and research findings in various scientific disciplines were covered with the "Science and Technology" class label.
- **Spiritual:** The news headlines which cover events, meetings, activities and the latest updates related to spiritual matters were labelled with the "Spiritual" class label.
- **Sports:** The news headlines which cover different sports events organized across the globe, sports schedules, and game results were labelled with the "Sports" class label.
- **Tourism:** The news headlines which cover updates, travel restrictions, transportation services, and festivals that happened over various tourist places across the globe were labelled with the "Tourism" class label.
- **Women Empowerment:** The news headlines which covers activities, events, developments, achievements and milestones that happened in women's life will be labelled with the "Women Empowerment" class label.

3.3.2 Annotation Procedure

After completion of the annotation schema design, we formed a team of three educated native Telugu speakers who were working in the academic field for the task of annotating extracted Telugu news headlines for classification. We gave clear guidelines to the annotators about the schema being used for annotating extracted Telugu news headlines. We asked the annotators to thoroughly understand the schema before starting the annotation process. Each annotator is required to annotate the Telugu news headlines with one of the class labels presented in the annotation schema.

3.3.3 Validation of Labels

Each Telugu news headline presented in the corpus has to be annotated by at least two annotators. Once all Telugu news headlines annotation has finished, we finalize the class labels for each Telugu news headline based on a common label assigned by both annotators. In case of disagreement with the label assignment for a news headline, a third annotator will annotate the Telugu news headline.

The most common label among the three annotators will be considered as a final annotation. Figure 2 presents sample news headlines and their corresponding labels, together with an estimated English translation of each news headline.

3.3.4 Assessment of Annotation quality

After completion of the validation of labels step, we evaluated the reliability of our annotation process. We conducted an inter-annotator agreement study on the annotated sentences to assess the reliability of the annotation process for the Telugu news headlines text classification task. In order to evaluate the efficacy of the annotation process, we choose Cohen's kappa metric to measure the quality of the annotation process. We got Cohen's kappa score as 0.95 which resembles a perfect agreement of the annotation process. Figure 1 presents the class label distribution of the proposed dataset for the Telugu News headlines classification dataset. Figure 2 displays a list of sample Telugu news headlines along with their labels and an approximation of their English translation.

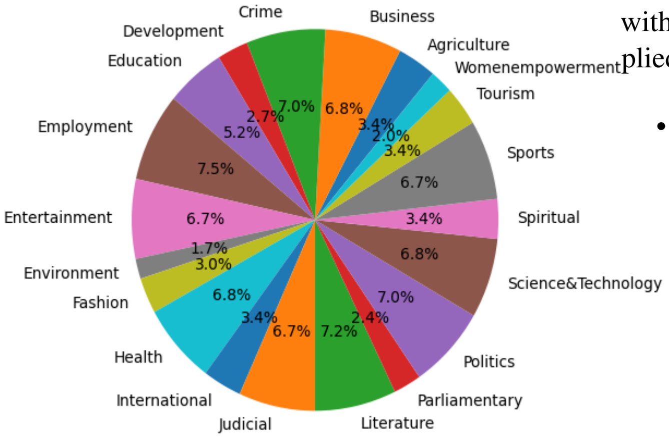


Figure 1: Class Label distribution in proposed dataset

4 Models

We evaluated multiple Machine Learning(ML) and Deep Learning(DL) models using different word embedding methods like TF-IDF, Word2Vec, and Glove with respect to the proposed dataset. In addition, we also fine-tuned different mono and multilingual pre-trained models in order to establish solid benchmark results. We started with ML models that are trained and evaluated using TF-IDF feature-based vectors. Next to this, a group of DL models were trained and evaluated with both Word2vec and Glove word embeddings. Finally, pre-trained transfer learning models were fine-tuned and evaluated with contextualized word embeddings. Multiple pre-trained transformer language models are adapted particularly for classification tasks on our dataset.

4.1 Machine learning models

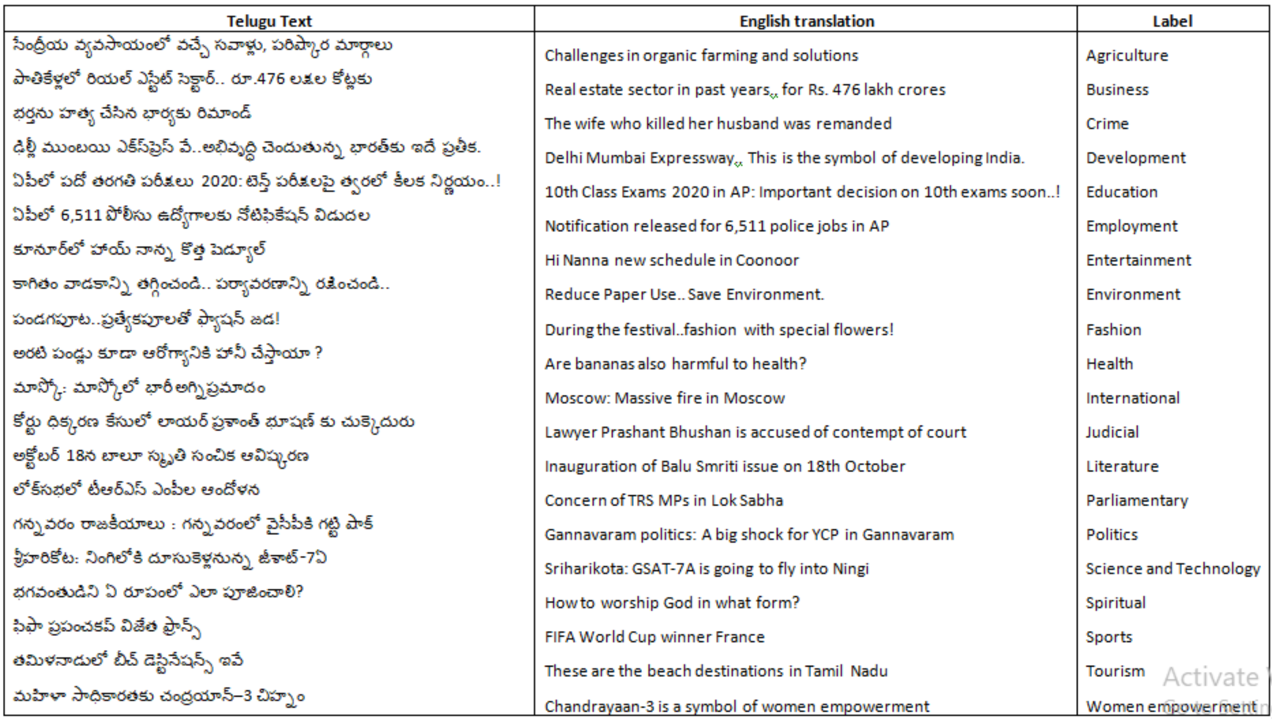
We trained five Machine Learning(ML) classification models, namely Naive Bayes(NB), Support Vector Machines(SVM), Random Forests(RF), Logistic Regression(LR) and Multi-Layer Perceptron. Each classifier is trained with TF-IDF-based feature vectors with optimal hyperparameter-configured values.

TF-IDF vectors are a method utilized to express textual data as numerical vectors, considering the frequency of a term within a specific document as well as its significance throughout a corpus of texts. The aforementioned format holds significant value in the context of text analysis and machine learning applications due to its ability to effectively capture the significance and distinctiveness of phrases within a given corpus of text. The ML models applied to the proposed corpus are described below.

- **Naive Bayes:** It is a simple probabilistic machine learning model that is used for classification. Its "Naive" name originates from its assumption of feature independence. The classifier uses Bayes' theorem and assumes all features are independent. Based on prior knowledge of class label-related events it determines class label probability.

- **SVM:** It stands for Support Vector Machines (SVM) and is a type of supervised machine learning algorithm that is mainly employed for problems involving classification and regression. SVM fundamentally seeks to identify the optimal hyperplane that effectively partitions data into distinct classes by maximizing the margin between these classes.

- **Random Forests:** It is a popular ensemble learning technique that is widely employed in the field of text classification. Its operational mechanism involves the construction of several decision trees for input features during the training phase. For an unknown sample, the class label is determined by selecting the



Telugu Text	English translation	Label
సేంద్రియ వ్యవసాయంలో వచ్చే సవాళ్లు, పరిష్కార మార్గాలు	Challenges in organic farming and solutions	Agriculture
పాతికేళ్లలో రియల్ ఎస్టేట్ సెక్టార్.. రూ.476 లక్షల కోట్లకు	Real estate sector in past years,, for Rs. 476 lakh crores	Business
భర్తను హత్య చేసిన భార్యకు రిమాండ్	The wife who killed her husband was remanded	Crime
ఢిల్లీ ముంబయి ఎక్స్‌ప్రెస్ వే..అభివృద్ధి చెందుతున్న భారత్‌కు ఇదే ప్రతీక.	Delhi Mumbai Expressway,, This is the symbol of developing India.	Development
ఏపీలో పదో తరగతి పరీక్షలు 2020: టెన్త్ పరీక్షలపై త్వరలో కీలక నిర్ణయం..!	10th Class Exams 2020 in AP: Important decision on 10th exams soon..!	Education
ఏపీలో 6,511 పోలీసు ఉద్యోగాలకు నోటిఫికేషన్ విడుదల	Notification released for 6,511 police jobs in AP	Employment
కూనూర్‌లో హాయ్ నాన్న కొత్త షెడ్యూల్	Hi Nanna new schedule in Coonoor	Entertainment
కాగితం వాడకాన్ని తగ్గించండి.. పర్యావరణాన్ని రక్షించండి..	Reduce Paper Use.. Save Environment.	Environment
పండగపూట..ప్రత్యేకపూలతో ఫ్యాషన్ జడ!	During the festival..fashion with special flowers!	Fashion
అరటి పండ్లు కూడా ఆరోగ్యానికి హానీ చేస్తాయా ?	Are bananas also harmful to health?	Health
మాస్కో: మాస్కోలో భారీ అగ్నిప్రమాదం	Moscow: Massive fire in Moscow	International
కోర్టు ధిక్కరణ కేసులో లాయర్ ప్రశాంత్ భూషణ్ కు చుక్కెదురు	Lawyer Prashant Bhushan is accused of contempt of court	Judicial
అక్టోబర్ 18న బాలూ స్మృతి సంచిక ఆవిష్కరణ	Inauguration of Balu Smriti issue on 18th October	Literature
లోక్‌సభలో టీఆర్ఎస్ ఎంపీల ఆందోళన	Concern of TRS MPs in Lok Sabha	Parliamentary
గన్నవరం రాజకీయాలు : గన్నవరంలో వైసీపీకి గట్టి షాక్	Gannavaram politics: A big shock for YCP in Gannavaram	Politics
శ్రీహరికోట: నింగిలోకి దూసుకెళ్లనున్న జీశాట్-7ఏ	Sriharikota: GSAT-7A is going to fly into Ningi	Science and Technology
భగవంతుడిని ఏ రూపంలో ఎలా పూజించాలి?	How to worship God in what form?	Spiritual
ఫిఫా ప్రపంచకప్ విజేత ఫ్రాన్స్	FIFA World Cup winner France	Sports
తమిళనాడులో బీచ్ డెస్టినేషన్స్ ఇవే	These are the beach destinations in Tamil Nadu	Tourism
మహిళా సాధికారతకు చంద్రయాన్–3 చిహ్నం	Chandrayaan-3 is a symbol of women empowerment	Women empowerment

Figure 2: Sample news headlines from proposed dataset

class that occurs most frequently among the trees.

- **Logistic Regression:** Logistic Regression is a statistical technique employed to estimate the likelihood of a categorical outcome, frequently applied in the context of text classification tasks. By employing a linear decision boundary, this system is capable of ascertaining the probability of a given text being associated with a specific class or category.

- **Multi-Layer Perceptron:** A Multi layer Perceptron (MLP) is a type of feed-forward artificial neural network with multiple layers of nodes that can capture complex patterns. An MLP can take input features, which are often taken from text data, and turn them into higher-level representations that can be used to put text into predefined groups.

4.2 Deep Learning Models

With respect to the proposed dataset, we trained and evaluated multiple DL models like Conv1D, GRU, BiGRU, LSTM, and BiLSTM with both Word2vec and Glove embeddings.

4.3 Pre-trained Language Models

Multilingual transformer-based language models are trained on text from multiple languages, enabling them to understand and generate content across different languages. Leveraging shared linguistic structures, these models, like mBERT and XLM, facilitate cross-lingual transfer learning, zero-shot learning, and improved performance for low-resource languages. We fine-tuned multilingual transformer-based language models on our proposed dataset for Telugu news headlines classification. These models were briefly discussed as follows:

- **M-BERT-Distil-40:** M-BERT-Distil-40 (Sanh et al., 2019) is a multi-lingual version of Distil-BERT. This multilingual language model was pre-trained over the top 100 languages during the pre-training phase, which also includes Telugu, and a list of 40 languages was used during the fine-tuning stage.

- **M-BERT-Base-ViT:** M-BERT-Base-ViT (Kenton and Toutanova, 2019) is a multi-lingual version of BERT-base. This model was trained over the top 100 languages during the pre-training stage which also includes

Telugu and is fine-tuned over a list of 69 languages during fine-tuning phase.

- **XLM-RoBERTa-large:**XLM-RoBERTa (Conneau et al., 2019) can be described as a multilingual variant of the RoBERTa model. This model has been pre-trained using 2.5 terabytes (TB) of filtered CommonCrawl data that contains 100 different languages. For our task, we fine-tuned **xlm-roberta-large** model.

- **MuRIL:** MuRIL (Khanuja et al., 2021) stands for Multilingual Representations for Indian Languages and was developed by google. This model is a variant of the BERT architecture, specifically tailored to accommodate the linguistic diversity observed in Indian languages. MuRIL has performed at the state-of-the-art level on multiple natural language comprehension benchmarks for the Indian languages. We fine tuned MuRIL-base version for our task.

- **Indic-BERT:**Indic-BERT (Kakwani et al., 2020) is a multilingual version of ALBERT pre-trained model. It has undergone pre-training on a corpus of 12 prominent languages spoken in India which also include Telugu. IndicBERT has significantly fewer parameters than other multilingual models (mBERT, XLM-R, etc.), yet its efficacy is comparable or superior to that of these models.

5 Experimental results

After completion of the annotation process, we finally had a dataset with 29744 news headlines. In order to advance with the experimental procedure, the dataset has been divided into train, validation, and test sets. Table 3 displays the distribution of class labels among the different divisions of the dataset. Table 4 offers a comparative examination of different models and their performances utilizing multiple text embedding techniques. Here, we describe the concise overview of experimental results as per different word embeddings.

- **TF-IDF**: Multiple ML techniques were tested with TF-IDF word embeddings. The SVM model has the highest precision, recall, accuracy, and F1-score, 59.17%, 57.71%, and 57.8%. MNB and LR models had similar performance with lesser precision, recall, accuracy, and F1-scores around 24%. Random Forests fared better with mid 50% metrics. MLP scored approximately 48% in all metrics.

- **Word2vec**:Multiple neural network models were employed with word2vec embeddings to classify news content. In the evaluations, the Conv1D model had the highest precision (63.74%), recall (63.27%), accuracy (63.27%), and F1-score (63.12%). GRU and LSTM architectures performed similarly, with GRU recording precision, recall, accuracy, and F1-scores of 52.93%, 51.36%, 51.36%, and 50.83%, and LSTM 50.34%, 48.73%, 48.74%, and 47.43%. BiGRU and BiLSTM have slightly lower measurements: 50.89% precision, 43.46% recall, 43.46% accuracy, and 42.38% F1-score for BiGRU, and 48.44%, 48.06%, 48.07%, and 47.28% for BiLSTM.

- **Glove:** Various neural network architectures were examined with Glove embeddings. The

Table 3: Class label distribution of Train, Validation and Test splits of proposed dataset.

Class Label	Train	Val	Test	Total
Agriculture	713	102	204	1019
Business	1408	201	402	2011
Crime	1447	206	414	2067
Development	566	81	162	809
Education	1088	155	312	1555
Employment	1609	230	460	2299
Entertainment	1400	200	400	2000
Environment	354	50	101	505
Fashion	630	90	180	900
Health	1413	202	404	2019
International	713	102	204	1019
Judicial	1401	200	400	2001
Literature	1491	213	426	2130
Parliamentary	508	73	145	726
Politics	1459	208	417	2084
Science&Technology	1404	200	402	2006
Spiritual	700	100	200	1000
Sports	1401	200	401	2002
Tourism	698	100	199	997
Women empowerment	417	59	119	595
Total	**20820**	**2972**	**5952**	**29744**

Table 4: Experimental results of various models over proposed Telugu20Newsgroup headlines dataset with respect to different word representations.

Embedding Model	Model	Precision	Recall	Accuracy	F1-Score
TF-IDF					
	SVM	**59.17**	**57.71**	**57.71**	**57.8**
	MNB	40.61	22.12	22.12	24.07
	LR	40.62	22.13	22.13	24.08
	RF	54.61	54.7	54.70	54.18
	MLP	47.98	47.71	47.69	47.64
Word2vec					
	Conv1D	**63.74**	**63.27**	**63.27**	**63.12**
	GRU	52.93	51.36	51.36	50.83
	BiGRU	50.89	43.46	43.46	42.38
	LSTM	51.34	48.73	48.74	47.43
	BiLSTM	48.44	48.06	48.07	47.28
Glove					
	Conv1D	74.99	74.42	74.42	74.42
	GRU	77.10	76.70	77.00	76.70
	BiGRU	74.48	74.12	74.12	74.19
	LSTM	77.10	76.76	76.76	76.75
	BiLSTM	**77.63**	**77.23**	**77.23**	**77.05**
Pre-train embeddings					
	M-BERT-Distil-40	80.1	79.56	79.56	79.42
	XLM-RoBERTa-large	**88.08**	87.56	87.56	87.64
	M-BERT-base	84.7	84.45	84.45	84.49
	MuRIL-base-cased	88.02	**87.95**	**87.95**	**87.92**
	Indic-BERT	71.59	70.78	70.78	70.76

Conv1D model had 74.99% precision, 74.42% recall, 74.42% accuracy, and F1-score 74.42%. Compared to Conv1D, the GRU model had 77.1% precision, 76.7% recall, 77% accuracy, and 76.7% F1-score. Precision, recall, accuracy, and F1-score were 74.48%, 74.12%, and 74.19 for the BiGRU model. The precision, recall, accuracy, and F1-scores of the LSTM and BiLSTM models were 77.1%, 77.63%, 76.76%,77.23%, and 77.05%, respectively. LSTM, BiLSTM, and GRU models had the highest precision and recall with glove embeddings.

- **Pre-trained word embeddings:**The XLM-RoBERTa-large and MuRIL-base-cased models performed well with pre-trained embeddings, scoring over 87% across all criteria. XLM-RoBERTa-large had 88.08% precision, 87.56% recall, 87.56% accuracy, and 87.64% F1-score. MuRIL-base-cased followed with 88.02% precision, 87.95% recall, accuracy, and F1-score 87.92%. The M-BERT-Base scores were equally strong at 84.5%. M-BERT-Distil-40 hit 80% in all measures. However, Indic-BERT lagged behind the others with precision and F1-score values around 71% and recall and accuracy just below 71%.

6 Results discussion

Figure 3 shows the confusion matrix of MURIL-base-cased model over T20News headlines dataset. Based on the this, we present a concise error analysis of the MURIL-base-cased model as follows.

- **Most common errors:** The instances of the "Science and Technology" class were misclassified as the "Business" class for a total of 22 times, indicating a significantly high frequency of misclassification. In a similar way, it was observed that the instances belonging to the "Business" class were falsely classified as instances of the "Science and Technology" class on 21 occasions. In addition, the news headlines pertaining to the "Science

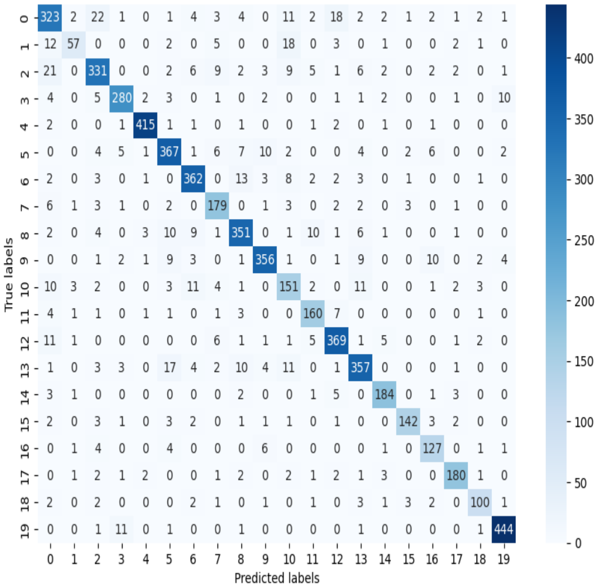


Figure 3: Confusion matrix for Telugu20Newsgroup headlines dataset. Here 0:ScienceandTechnology, 1: Environment, 2:Business, 3:Education, 4:Literature, 5:Politics, 6:Sports, 7:Agriculture, 8:Entertainment, 9:Judicial, 10:International, 11:Fashion, 12:Health, 13:Crime, 14:Spiritual, 15:Development, 16: Parliamentary, 17:Tourism, 18:Women-empowerment, 19:Employment

and Technology" class were incorrectly categorized as "Health" class on a total of 18 occasions. Instances of the "Environment" class were wrongly classified as the "International" class in equal number of proportions.

- **Potential issues:** The "Science and Technology" class exhibits misclassification that are distributed among many classes, notably including Business, Health, International, and Agriculture, which have substantial values. This observation implies that there may be a presence of overlapping features between "Science and Technology" and other classes such as Business, Health, and International can lead to misclassification. Likewise, the "Environment" category exhibits a significant occurrence of misclassification concerning the "International" category, amounting to a total of 18 cases.

- **High Accuracy Classes:** Classes such as "Education", "Literature", "Sports", "Judicial", "Spiritual", "Development", "Tourism", and "Employment" have higher F1-score when compared with other classes and overall model F1-score. This scenario indicates that, these classes instances were misclassified less in number and the model is performing well for these classes.

- **Improvement Areas:** The classes like "Environment", "International", "Science and Technology","Business", and "Women empowerment" instances had higher percentage of misclassified instances with respect to their total number of class instances, thus the model has to improve. The classes "Business" and "Science and Technology" instances were misclassified from each other. To reduce this, we need more distinguishing features to make accurate classification. The classes "Environment", "International", and "Women Empowerment" need more training samples to improve classification accuracy.

7 Conclusion

This paper presents the development of an extensive dataset for the purpose of text classification of Telugu news headlines.Data was gathered from a diverse range of online news publishing platforms, afterwards subjected to pre-processing techniques, and ultimately utilized to construct an annotated corpus. Various models were utilized in our study to classify Telugu news headlines text, employing multiple word embeddings methodologies. In this study, we have successfully constructed baselines on proposed dataset for the purpose of text classification. This has allowed us to create and evaluate a reliable dataset specifically designed for a low-resource language, such as Telugu.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Dheeru Dua and Casey Graff. 2017. UCI machine learning repository.

A Kanaka Durga and A Govardhan. 2011. Ontology based text categorization-telugu document. *International Journal of Scientific and Engineering Research*, 2(9):1–4.

Veerraju Gampala, Jaideep Vallapuneni, Pavan Kumar Ande, Ravindra Kumar Indurthi, and Nichenametla

Rajesh. 2021. Comparative study on telugu text classification using machine learning and deep learning models. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1393–1398. IEEE.

Rama Rohit Reddy Gangula and Radhika Mamidi. 2018. Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Sunil Gundapu, IIIT KCIS, and Radhika Mamidi. 2020. Multichannel lstm-cnn for telugu technical domain identification. In *17th International Conference on Natural Language Processing*, page 11.

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

Rohan Katari and Madhu Bala Myneni. 2020. A survey on news classification techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–5. IEEE.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Siva Subrahamanyam Varma Kusampudi, Anudeep Chaluvadi, and Radhika Mamidi. 2021. Corpus creation and language identification in low-resource code-mixed telugu-english text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 744–752.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.

Sandeep Sricharan Mukku and Radhika Mamidi. 2017. Actsa: Annotated corpus for telugu sentiment analysis. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 54–58.

Kavi Narayana Murthy. 2003. Automatic categorization of telugu news articles. *Department of Computer and Information Sciences*.

Vishnu G Murthy, B Vishnu Vardhan, K Sarangam, and P Vijay Pal Reddy. 2013. A comparative study on term weighting methods for automated telugu text categorization with effective classifiers. *International Journal of Data Mining & Knowledge Management Process*, 3(6):95.

Swapna Narala, B Padmaja Rani, and K Ramakrishna. 2017. Telugu text categorization using language models. *Global journal of computer science and technology*.

Sreekavitha Parupalli, Vijjini Anvesh Rao, and Radhika Mamidi. 2018. Bcsat: A benchmark corpus for sentiment analysis in telugu using word-level annotations. In *Proceedings of ACL 2018, Student Research Workshop*, pages 99–104.

Yashwanth Reddy Regatte, Rama Rohit Reddy Gangula, and Radhika Mamidi. 2020. Dataset creation and evaluation of aspect based sentiment analysis in telugu, a low resource language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5017–5024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Vukyam Sri Sravya, Sachin Kumar, and KP Soman. 2022. Text categorization of telugu news headlines. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6. IEEE.

D Naga Sudha et al. 2021. Semi supervised multi text classifications for telugu documents. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(12):644–648.

M Narayana Swamy, M Hanumanthappa, and NM Jyothi. 2014. Indian language text representation and categorization using supervised learning algorithm. In *2014 International Conference on Intelligent Computing Applications*, pages 406–410. IEEE.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022. N24news: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 649–657.