A Dataset for Detecting Humor in Telugu Social Media Text

Sriphani Vardhan Bellamkonda
National Institute of Technology Warangal
Telangana, India
sriphani345v@gmail.com

Maithili Lohakare
Pandit Deendayal Energy University
Gandhinagar, Gujarat, India
maithililohakare@gmail.com

Shaswat P Patel
Netaji Subhas University of Technology
New Delhi, India
shaswat178@gmail.com

Abstract

Increased use of online social media sites has given rise to tremendous amounts of user generated data. Social media sites have become a platform where users express and voice their opinions in a real-time environment. Social media sites such as Twitter limit the number of characters used to express a thought in a tweet, leading to increased use of creative, humorous and confusing language in order to convey the message. Due to this, automatic humor detection has become a difficult task, especially for low-resource languages such as the Dravidian languages. Humor detection has been a well studied area for resource rich languages due to the availability of rich and accurate data. In this paper, we have attempted to solve this issue by working on low-resource languages, such as, Telugu, a Dravidian language, by collecting and annotating Telugu tweets and performing automatic humor detection on the collected data. We experimented on the corpus using various transformer models such as Multilingual BERT, Multilingual DistillBERT and XLM-RoBERTa to establish a baseline classification system. We concluded that XLM-RoBERTa was the best-performing model and it achieved an F1-score of 0.82 with 81.5% accuracy.

1 Introduction

The use of social media sites has increased exponentially over the decade giving rise to vast amount of user generated content. Social media sites offer the ability to reach large number of users in real time which enable users to share their experiences easily. The content usually consists of creative and figurative use of languages such as humor, insults, sarcasm and irony. In the past couple of years, research in these linguistic elements has increased tremendously due to requirements in academia as well as in organizations.

Natural language processing(NLP) has evolved significantly leading to improvements in most of the fundamental tasks like Named-Entity recognition, sentiment analysis, etc (Singh etal., 2021). While the advancement is not only attributed to improvements in architecture of models but also due to the increased availability of data. Plethora of work exists for resource rich languages such as English (VanHee etal., 2018; A. and Sonawane, 2016; Patel etal., 2022). However, the same cannot be said for low-resource languages originating from the Indian subcontinent such as the Dravidian languages. Telugu is one of the four major Dravidian languages that stem from India, it is spoken by more than 75 million people (top, 2005). Hence, it is vital to establish a baseline system for automatic humor detection in Telugu language.

In this paper, we explore the task of humor detection, one of the critical elements of a natural language (Kruger, 1996). Humor is subtle and yet plays a significant part in our linguistic and social lives (Martin, 2007). The primary challenge in working with humor detection is the subjective nature of humor and capturing it in higher order is a challenge in NLP (deOliveira and Rodrigo, 2015). Yet, a large amount of research work has been carried out on English tweets and has achieved significant results.

One of the major challenges in building any social media analysis model in low-resource languages is the unavailability of high quality annotated data for conducting various experiments. In this paper, we address this issue by collecting Telugu tweets data by scraping Twitter and performing annotations on it. Furthermore, the dataset we collected is publicly available [1]. Below is an example of humorous tweet in Telugu:

[1] https://github.com/shaswa123/telugu_humour_dataset

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 9 - 14
May 26, 2022 ©2022 Association for Computational Linguistics

నవ్వు రాకపోయినా నవ్వేవాలని ఏమంటారో కానీ ?
What do you call those people who laugh even when laughter is not induced.

Context: This tweet intends to mock the judges of a Telugu comedy show who laugh a lot.

Additionally, we trained three multilingual transformer-based models, namely: Multilingual BERT, Multilingual DistilBERT, and XLM-RoBERTa and compared their performances to establish baseline classification system for humor detection in Telugu.

The rest of the paper is organized into related works (Section 2), detailed description of the dataset (Section 3), brief description of the methodology used(Section 4), analysis of results (Section 5), and finally conclusion (Section 6).

2 Related works

Plenty of work exists on social media text analysis including humor detection. Various works related to humor detection exist in English language, such as statistical and N-gram analysis (Rayz and Mazlack, 2004), Regression Trees (Purandare and Litman, 2006), Word2Vec combined with K-NN Human Centric Features (Yang etal., 2015), Convolutional Neural Networks (Chen and Soo, 2018) and transformer models (Weller and Seppi, 2019).

Previous work related to humor detection in Hindi-English mixed language dataset consists of scraping Hindi-English tweets and building N-grams, Bag-of-words, LSTM, Bi-directional LSTM, and Attention Based Bi-directional LSTM (Khandelwal etal., 2018; Agarwal etal., 2021; Sane etal., 2019).

Related works for humor detection in Telugu language is scarce. Vaishnavi et. al (Pamulapati and Mamidi, 2021) proposed conversational data in Telugu language for humor detection and experimented with TextGCN, FastText, Multilingual BERT, MuRIL, Indic-BERT, and Multilingual DistilBERT models. Automatic humor detection in Telugu for Twitter data is an under-explored area. According to our knowledge, this paper is the first attempt at building a novel telugu dataset and then proceeding with experimenting on various classifiers for humor detection task.

Category	Tweets	Words
Humorous	458	5477
Non-Humorous	1918	18098
Other	273	4213

Table1: Telugu Twitter humor corpus statistics

3 Dataset

3.1 Corpus Creation

For the creation of our dataset, we have scraped tweets from Twitter by filtering specific tags. For collecting humorous tweets, we searched using the tags such as humour, humor, funny, telugu-jokes. Additionally, we also searched using telugu hashtags such a s తమాషా and నవ్వు. In total we collected 1649 tweets using the above-mentioned tags. For non-humorous tweets we searched using tags such as news, sports, cooking, cinema etc. in order to include tweets from various domains in our dataset. Additional 1000 tweets were collected with these tags. The statistics of the resulting dataset is shown in Table 1. All these tweets were then annotated via two human annotators.

3.2 Humor annotation

The annotation of our tweets was done by two native multilingual Telugu speakers. Around 50 human hours was spent in tagging tweets into 3 categories: 0 for non-humorous, 1 for humorous, and n for tweets which do not have enough context to be considered informative or whose body was repeated. A tweet was considered humorous if it consisted sarcasm, irony, comedy, mockery, comment, or insult. Tweets which were just stating facts, general speech, quotes, or did not have any sort of amusement were considered non-humorous. These humor specifications were taken from (Khandelwal etal., 2018). The 2649 tweets were fairly split between the 2 annotators. Below are few examples of tweets from our corpus:

1. అరటి ఆకులో 66 కూరలు ఉన్నా ఆవకాయ లేదా అని అదిగెవాడే మన తెలుగు వాడు
   Even though there are 66 curries in a banana leaf, the one asking for mongo pickle is the true telugu person.
   Explanation: This tweet is classified as humorous as it wittily describes the telugu people's liking for mango pickle. Obviously, not all have this preference, but it is considered as a popular liked dish in the region.

2. మన సినిమా ల కి ఆడియన్సే రారు ..ఇక ' ఆస్కార్ ' రాటానికి అస్కారం ఎక్కడుంటుంది .
Our movies aren't even watched by our own audience, so for getting oscars we have no scope.
Explanation: This tweet is classified as humorous as it is satirical in nature and the user makes a clever word play between oscar, prestigious award given to movies, and "askaram", a telugu word which means having scope for.

3. సమతామూర్తి విగ్రహాన్ని దర్శించుకున్న కేంద్ర మంత్రి రాజ్‌నాథ్ సింగ్
Union Minister Rajnath Singh visited the Samathamurthy statue
Explanation: This tweet is classified as non humorous as it just states a fact and doesn't contain any comedy, satire, or amusement.

4. ఎయిర్‌టెల్ ఇంటర్నెట్ డౌన్... ఫన్నీ మీమ్‌లతో పిచ్చెక్కిచ్చిన నెటిజన్లు!
Airtel internet down ... insane netizens with funny memes!
Explanation: This tweet is marked as other and will be discarded as it isn't inherently humorous but refers to some memes which might be.

3.3 Inter Annotator Agreement

We used the Cohens Kappa coefficient for calculating the Inter Annotator Agreement as only 2 annotators were involved. The annotators are Telugu-native speakers whose second language is English. We extracted a set of 100 tweets and provided it to the annotators to measure their agreement score. The first 50 were sampled from the 1649 tweets intended to be humorous and the next 50 were sampled from the 1000 tweets intended to be non-humorous. We obtained a Kappa score of 0.84, implying that the annotation is of high quality.

4 Methodology

4.1 Preprocessing

In order to convert the humorous and non-humorous data to a favorable format for performing humor detection task, several changes were made to the collected tweets. Preprocessing steps such as removal of emojis, hashtags and URL links were implemented. In the tweets, URL links do not add any information to the data and were hence, removed. Additionally, hashtags were removed so that the models would classify effectively without any bias towards the tags used for filtering the tweets while scraping.

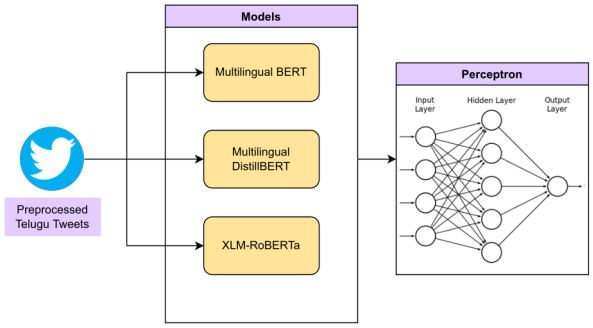

Figure1: Outline of Methodology

4.2 Outline of Methodology

The preprocessed data was then used by the transformer-based models in order to learn distinguishable features to classify tweets into humorous and non-humorous. The input text was initially tokenzied using Subword or Sentencepiece tokenzation method (Devlin etal., 2019; Liu etal., 2019). The tokenized tweet in addition with segment ID and attention mask is fed to the transformer models to generate meaningful vectors summarizing the context of the tweet. The fix size vector is then classified using a single layered perceptron as shown Figure 1.

5 Modelling

We have trained three transformer models on our corpus for humor detection task. Transformer models with their improved architecture have achieved state-of-the-art results in numerous benchmark datasets including text classification tasks (Young etal., 2018; Alam etal., 2021). Hence, we decided to train these models to establish a suitable baseline for future work.

5.1 Transformers

We have used BERT and its variations such as Multilingual BERT, Multilingual Distill BERT and XML-RoBERTa, which are all transformer-based models, as our primary models for performing the humor detection task. Transformers are a set of deep learning models which use attention-based mechanism and are used for transforming one sequence into another using encoders and decoders. The architecture of a Transformer model consists

of the input being passed through an encoder which has two parts: a multi-headed self attention layer and a feed-forward network. This information from the encoder is then presented as output into the decoder which includes the same above-mentioned parts but with an additional masked attention step. Lastly, it is transformed through a softmax layer into the output. The Transformer's self-attention layers are greatly attributed for its success. And hence, we chose to use the Transformer based models for their efficiency at recognizing and attending to the relevant words in sentences and paragraphs which would help classify the tweets with more accuracy and precision.

5.2 Multilingual BERT

Bidirectional Encoder Representations from Transformers or BERT is a transformer-based architecture (Devlin etal., 2019) that has greatly outperformed previous models like RNN-based models in various benchmark datasets. This is due to the ability of the model to capture latent information from text successfully into a fixed sized vector. This is mainly attributed to the following two tasks on which the BERT model was trained on:

1. Masked Language Model(MLM): From the given input sequence, 15% of tokens are randomly chosen and replaced with [MASK] tokens. The objective of this task is to correctly predict the masked tokens.

2. Next sentence prediction(NSP): From the given input segments, the task is to predict whether the input segments follow each other in the original text.

Multilingual BERT is trained on Wikipedia data consisting of 104 different languages (Pires etal., 2019). It has shown better accuracy as compared to BERT for NLP tasks involving machine translation and tasks dealing with multiple languages. Moreover, out of 104 languages, Telugu was one of the languages the multilingual BERT was pre-trained on. Hence, it became crucial to test our corpus by training this model on it.

5.3 Multilingual DistilBERT

Multilingual DistilBERT is the distilled version of Multilingual BERT (Sanh etal., 2019). It is also trained on text belonging to 104 different languages including Telugu and on the same Wikipedia dataset as Multilingual BERT. It consists of 134 million parameters making it on average twice as faster as multilingual BERT, hence, making it cheaper to train and convenient to test on our corpus.

5.4 XLM-RoBERTa

XLM-RoBERTa is a multilingual version of RoBERTa (Conneau etal., 2020). RoBERTa is a transformer based model which also happens to be an improved version of BERT. The following modifications were implemented on training of BERT to improve it's performance:

1. Model was trained on bigger batches and for more epochs.

2. Removing next sentence prediction task from the training objective.

3. Longer sequences were considered for training the model.

4. Changing the masking pattern dynamically for the training data.

The XLM-RoBERTa was pre-trained on 100 different languages using over 2.5TB of filtered CommonCrawl data (Conneau etal., 2020). Moreover, the vocabulary size was also significantly larger in comparison to multilingual BERT. These modifications to BERT have made the XLM-RoBERTa a more robust model and made it outperform multilingual BERT significantly in most of the multilingual NLP tasks. Therefore, XLM-RoBERTa was the best choice for our task.

6 Results

The corpus was split into 80% train and 20% test. We have downsampled the non-humorous tweets to match the number of humorous tweets. At the end, we have 732 training examples and 184 test examples. We have considered macro F1-score as our metric to select the best model out of the three models trained on our corpus. XLM-BERT has outperformed other models with a F1-score of 0.82 and an accuracy of 81.5% as shown in Table 2. XLM-RoBERTa has shown a significant improvement over multilingual BERT in various multilingual tasks (Conneau etal., 2020). This is mainly due to the increase in vocabulary size and in the amount of training data over multilingual BERT.

Model	Accuracy	F1-score
Multilingual BERT	81.5	0.81
Multilingual DistilBERT	73.4	0.73
XLM-RoBERTa	81.5	0.82

Table2: Results of various models trained and tested on our corpus.

7 Conclusion and Future work

In this paper, we introduced the Telugu Twitter Humor Dataset and have addressed the need to annotate low-resource languages and create datasets in languages such as Telugu. We have also described our data collection and annotation process. Additionally, we have trained multiple transformer-based models, namely, Multilingual BERT, Multilingual DistillBERT, and XML-RoBERTa, to perform automatic humor detection on our collected dataset. The performance of these models is compared and a baseline for classification of humorous and non-humorous Telugu tweets is established in this paper. Out of the above-mentioned models used for training our data, XLM-RoBERTa performed the best with a F1-score of 0.82 and an accuracy of 81.5%. We would like to expand this work in the future by incorporating the information detected from emojis into the classifiers and by making a multi-modal humor detection classifier as a large number of tweets which were discarded had images present in them too.

Acknowledgements

We thank our anonymous reviewers for providing their valuable feedbacks. All the opinions, conclusions and findings presented in this material are those of the authors only and do not reflect the views of their graduate schools or employing organizations. We would also like to thank our annotators, Mrs. Bellamkonda Aruna and Mr. Bellamkonda Kiran Kumar for their effort in annotating the data and review.

References

2005. Top 30 languages by number of native speakers.

Vishal A. and S.S. Sonawane. 2016. Sentiment analysis of twitter data: A survey of techniques. *International Journal of Computer Applications*, 139(11):515.

Kaustubh Agarwal, , and RhythmNarula and. 2021. Humor generation and detection in code-mixed hindi-english. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*. INCOMA Ltd. Shoumen, BULGARIA.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and ShammurAbsar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Luke deOliveira and AlfredoLáinez Rodrigo. 2015. Humor detection in yelp reviews.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ankush Khandelwal, Sahil Swami, SyedS. Akhtar, and Manish Shrivastava. 2018. Humor detection in english-hindi code-mixed social media content : Corpus and baseline system.

Arnold Kruger. 1996. The nature of humor in human nature: Cross-cultural commonalities. *Counselling Psychology Quarterly*, 9(3):235–241.

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

RodA. Martin. 2007. *The Psychology of Humor*. Elsevier.

Vaishnavi Pamulapati and Radhika Mamidi. 2021. Developing conversational data and detection of conversational humor in Telugu. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 12–19, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Shaswat Patel, Binil Shah, and Preeti Kaur. 2022. Leveraging user comments in tweets for rumor detection. In *International Conference on Innovative Computing and Communications*, pages 87–99, Singapore. Springer Singapore.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S*. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215, Sydney, Australia. Association for Computational Linguistics.

JuliaTaylor Rayz and LawrenceJ. Mazlack. 2004. Computationally recognizing wordplay in jokes.

SushmithaReddy Sane, Suraj Tripathi, KoushikReddy Sane, and Radhika Mamidi. 2019. Deep learning techniques for humor detection in Hindi-English code-mixed tweets. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–61, Minneapolis, USA. Association for Computational Linguistics.

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

Shaanya Singh, Maithili Lohakare, Keval Sayar, and Shivi Sharma. 2021. Recnn: A deep neural network based recommendation system. In *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, pages 1–5.

Cynthia VanHee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing.