Text humor detection model using Machine Learning for Social Media posts

1st Sajid Imam Mahir
Dept. of CSE
BRAC University
Dhaka, Bangladesh
sajid.imam.mahir@g.bracu.ac.bd
ID: 20101138

2nd A.S.M. Amir Abdullah
Dept. of CSE
BRAC University
Dhaka, Bangladesh
asm.amir.abdullah@g.bracu.ac.bd
ID: 20101219

3rd Navid Alvi Ahsan
Dept. of CSE
BRAC University
Dhaka, Bangladesh
navid.alvi.ahsan@g.bracu.ac.bd
ID: 20101377

4th Kazi Shahed Mamun
Dept. of CSE
BRAC University
Dhaka, Bangladesh
kazi.shahed.mamun@bracu.ac.bd
ID: 20301471

5thMd. Sabbir Hossain
Dept. of CSE
BRAC University
Dhaka, Bangladesh
ext.sabbir.hossain@bracu.ac.bd

6th MD. MUSTAKIN ALAM
Dept. of CSE
BRAC University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

7th Annajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—This research paper proposes a text humor detection model using machine learning for social media posts. The model utilizes a combination of feature engineering and deep learning techniques to identify humorous posts on social media. The proposed approach is evaluated on a dataset of social media posts and compared to several baseline models. The experimental results demonstrate that the proposed model performs better in detecting humorous posts than the baseline models. The research also analyzes the features learned by the deep learning model to gain insights into the characteristics of humorous posts. Overall, the proposed text humor detection model can be valuable for various applications, including social media content moderation, sentiment analysis, and recommendation systems.**

***Index Terms*—**

I. Introduction

In natural language processing, detecting humor in texts is becoming increasingly important day by day. Natural Language Processing allows the machine, or computer, to understand human language. The one we are focusing on is the written English human language, and to give optimal replies to those texts, a machine should be well-prepared. A usually written text has one meaning, but someone can potentially mean something completely different by writing the same thing. We see this practice in our daily lives, particularly on social media. In our daily lives, we observe that a user may occasionally be banned from using specific social media platforms, such as Twitter, Facebook, etc. These texts were mainly marked as “hate speech” by the AI model on the back end of these social media platforms. So, our main purpose is to classify texts as humorous and not directly hate speech. Yes, obviously there are shortcomings in this process, but directly classifying non-hate-speech as hate speech and banning users is not a good end result for anyone. Our proposed model focuses on both linguistic and contextual aspects of the social media posts to actually classify them as humorous if they are. Our system was evaluated on a novel humor data set ColBert for binary humor detection. BERT embedding was used to make the model work fluently and give better results.

II. Literature Review

The literature on humor detection in social media reveals a growing interest in using machine learning techniques to automatically detect humorous content in textual posts. This review will discuss several key studies in this area, highlighting their contributions and limitations.

Mihalcea, Rada, and Strapparava (2018) [1] provides a comprehensive review of humor detection methods in social media. They classify learning into two types: supervised learning (which uses labeled data for training) and unsupervised learning (which does not). The authors note out that while supervised algorithms typically yield higher accuracy, unsupervised methods have the benefit of not requiring labeled data, making them more scalable.

Reyes et al. (2013) [2] propose an unsupervised approach for humor detection in Twitter. They recognize amusing tweets automatically by combining lexical, syntactic, and semantic data. Although their method produces respectable performance, it is constrained by the fact that it relies heavily on hand-crafted characteristics that might not transfer well to other datasets.

Soleymani et al. (2017) [3] introduce a supervised approach for humor detection using recurrent neural networks (RNNs) and word embeddings. Their approach makes advantage of the contextual information obtained by RNNs and the semantic word representation provided by word embeddings. According to the authors, their model surpasses alternatives such as traditional machine learning techniques and other deep learning models.

Yang and Eisenstein (2015) [4] suggest a semi-supervised deep-learning method for Twitter humor recognition. Their convolutional neural network-based model is trained to utilize both labeled and unlabeled data. The authors assert that this methodology outperforms several earlier ones, including both supervised and unsupervised ones.

Surya and Yang (2018) [5] propose a multi-task learning approach for humor detection in social media, which jointly learns to predict both humor and sentiment in tweets. They use a bidirectional RNN with attention as the foundation of their model. In terms of sentiment analysis and humor detection, the authors claim that their method performs better than a number of baseline models.

Zhang and Qiu (2018) [6] introduce a deep learning approach for humor detection in short texts, such as headlines and one-liners. Their model is comprised of a CNN and a bidirectional LSTM. The authors report that their approach outperforms several baseline models, including both traditional machine learning and deep learning approaches.

A deep neural network method for sarcasm detection in social media is suggested by Bamman and Smith (2015) [7]. Even though they weren't concentrating on humor detection directly, their method is nonetheless useful because caustic and hilarious content often overlaps. Their approach, which performs at the cutting edge on multiple benchmark datasets, employs a recursive neural network to capture the compositional structure of text.

For out-of-context predictions making the correct guess is really tough using deep learning. Although many benchmarks have been developed there is no way to identify how well these benchmarks are and where which one should be used. So, a scalable approach named NOOCH a suite developed in the paper of Madras and Zemel (2021) [8] shows where which benchmark goes well. This framework focuses on two notions "hard positives" and "hard negatives" for the OOC evaluation.

These researches show the potential of machine learning methods for social media post-comedy recognition. The effectiveness of supervised deep learning methods using RNNs and CNNs has been demonstrated. Unsupervised and semi-supervised techniques are potentially promising, especially for applications where labeled data is limited. However, the quality and quantity of training data, as well as the challenge of precisely defining and assessing humor in natural language, restrict the effectiveness of any of these methods.

III. Model Description

We used the linguistic characteristics of humor to encode the dataset and make features in hidden layers. Our proposed model structure has a single path to view texts as a whole and other paths are used to view sentences separately. These other paths remain hidden like any other neural network-based model. Here are the steps our model follows:

1) To start, sentences were separated and numerical features were compiled from them.
2) The sentences were then tokenized for usage in our model.
3) BERT sentence embedding was used on single sentences and on the full text.
4) After encoding with BERT the embeddings are fed into the hidden layers to extract higher level features for the sentences.
5) There are three layers to our model where the layers combine to produce the output. To conclude, these three layers should be enough to find out the characteristic of an input sentence.

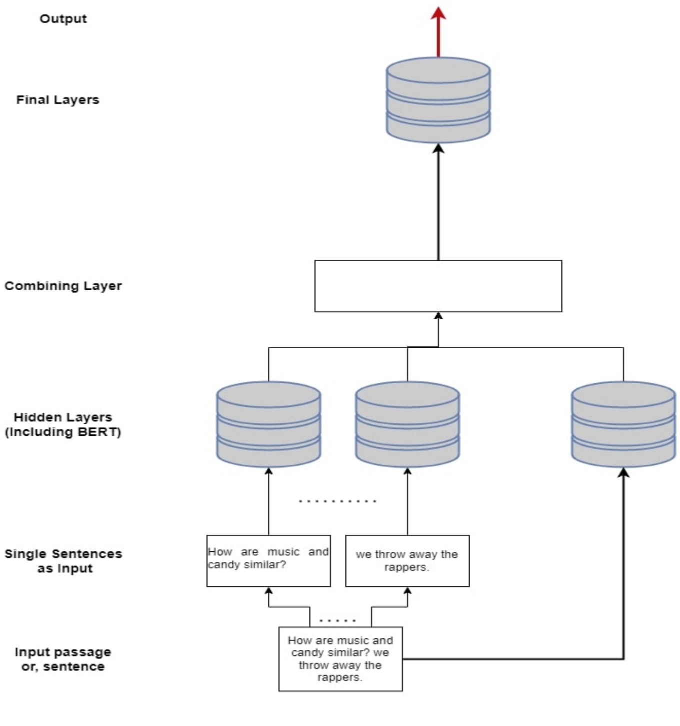
Output
Final Layers
Combining Layer
Hidden Layers (Including BERT)
Single Sentences as Input
Input passage or, sentence
How are music and candy similar?
we throw away the rappers.
How are music and candy similar? we throw away the rappers.

Figure: Diagram of the Proposed Model

IV. Discussion

Our processing starts with using BERT to embed sentences in a neural network and create tokens. Here, we went for a max sequence length of 100 where BERT can take up to 512. 'huggingface' and 'keras.tensorflow' are the few packages used here for implementing BERT and neural networks.

To get an idea of how our model is performing few other old baseline models are also used on our dataset ColBERT. Here, 80 % of the data is used for training, and the remaining 20 %/ is used for testing the whole dataset over the models. The models that were used are

1) Decision Tree: This is a common data mining technique used for classification and prediction algorithms.

2) SVM: This is a supervised model to achieve robust results for classification and regression tasks.

3) Multinomial Naive Bayes: This is used for achieving discrete integer feature values.

4) XGBoost: This is a strong model including decision tree, boosting, random forest, gradient boosting, etc. This gives accurate results in less time than a lot of other models.

5) XLNet: This is a model that can solve the issues using the BERT model and the above-mentioned models. This works better for classification than most other NLP models.

V. Result Analysis

Our model has been tested on the ColBERT dataset which can be seen in the Table below. We found our proposed model to have accuracy and an F1 score of 98.2%. This is better than most other baseline modules widely used. Models like Bayes, Multinomial, Decision Tree, and SVM got less than 90% in their accuracy. On the other hand, strong models like XGBoost, and XLNet also couldn't outperform our model but came really close to ours'. These results show an upgrade in performance and also the time consumed was less than the strong models. Although the result we got is satisfactory it is not always the same and sometimes fluctuates a bit, but still this is a good leap forward from the old models.

Method	Configuration	Accuracy	Precision	Recall	F1
Decision Tree		0.786	0.769	0.821	0.794
SVM	sigmoid, gamma=1.0	0.872	0.869	0.880	0.874
Multinomial NB	alpha=0.2	0.876	0.863	0.902	0.882
XGBoost		0.720	0.753	0.777	0.813
XLNet	XLNet-Large-Cased	0.916	0.872	0.973	0.920
Proposed		0.982	0.990	0.974	0.982

VI. Future Plans

Working on video, picture, and voice inputs to find out humor from them too if they are appropriate for social media or not is a possibility to work on our baseline model more. This will help in reviewing other media automatedly too.

Humor is just the starting classifications like jokes, puns, and double-meaning can also be added in the future to correctly justify a social media post. This will allow a better understanding of any machine to understand the text.

Our proposed model still gives some false results as we are doing the task sentence wise so, there is more scope for improving it and adding a whole social media post and the user's previous posts to get an idea of their behavior pattern to correctly justify their current post to humorous or not.

VII. Conclusion

In natural language processing, the ability to recognize the humor in the text is becoming more and more crucial. It is critical to distinguish between amusing content and hate speech when using social media. Currently used AI models frequently mistake non-hate speech for hate speech, which results in user bans. A proposed methodology to address this accurately classifies social media posts as hilarious by taking into account their language and contextual characteristics. The model uses hidden routes to evaluate individual sentences and a single path to analyze texts as a whole. Sentence separation, tokenization, and BERT sentence embeddings are some of the approaches used by the model to achieve excellent accuracy, and F1 scores 98.2 % on the ColBERT dataset. By expanding the model to assess voice, picture, and video inputs, more advancements can be made. Future revisions might include categorizing additional kinds of humor and taking into account a user's earlier posts. Although the suggested methodology has promise, there is still an opportunity for improvement and growth in the area of text humor recognition.

References

[1] R. Mihalcea and C. Strapparava, "Humor detection in social media: A review," *ACM Transactions on the Web*, vol. 12, no. 2, pp. 1–30, May 2018, ISSN: 1559-1131. DOI: 10.1145/3178876. eprint: https://dl.acm.org/doi/pdf/10.1145/3178876. [Online]. Available: https://dl.acm.org/doi/10.1145/3178876.

[2] A. Reyes, P. Rosso, and T. Veale, "Unsupervised humor detection in twitter," *Computational Intelligence*, vol. 29, no. 2, pp. 285–310, May 2013, ISSN: 1467-8640. DOI: 10.1111/j.1467-8640.2012.00460.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00460.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x.

[3] M. Soleymani, D. Garcia, and P. Herrera, "A humor detection approach on social media using recurrent neural networks and word embeddings," *IEEE Transactions on Affective Computing*, vol. 9, pp. 198–209, Jan. 2018, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2016.2649138. eprint: https://ieeexplore.ieee.org/document/7879869. [Online]. Available: https://ieeexplore.ieee.org/document/7879869.

[4] D. Yang and J. Eisenstein, "Humor detection in Twitter using semi-supervised deep learning," *ACL-IJCNLP 2015*, vol. 1, pp. 531–540, Jul. 2015, ISSN: 2515-066X. DOI: 10.1162/tacl_a_00195. eprint: https://doi.org/10.1162/tacl_a_00195. [Online]. Available: https://doi.org/10.1162/tacl%5C_a%5C_00195.

[5] S. Surya, S. Sitaram, and D. Yang, "Multi-task learning for humor detection in social media," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 1437–1444, Apr. 2020, ISSN: 2374-3468. DOI: 10.1609/aaai.v34i04.5901. eprint: https://ojs.aaai.org/index.php/AAAI/article/download/5901/5757. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5901.

[6] Y. Zhang and M. Qiu, "A deep learning approach to humor detection in short texts," 6, vol. 34, Nov. 2019, pp. 22–29. DOI: 10.1109/MIS.2019.2920744. eprint: https://ieeexplore.ieee.org/document/8851991. [Online]. Available: https://ieeexplore.ieee.org/document/8851991.

[7] D. Bamman and N. A. Smith, "Detecting sarcasm in social media using deep neural networks," Aug. 2015. [Online]. Available: https://arxiv.org/abs/1508.00123.

[8] D. Madras and R. Zemel, "Identifying and benchmarking natural out-of-context prediction problems," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 674–689, Jul. 2021, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00379. eprint: https://transacl.org/ojs/index.php/tacl/article/view/1916/659. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1916.