CEUR-WS.org/Vol-2943/haha_paper1.pdf

TECHSSN at HAHA @ IberLEF 2021: Humor Detection and Funniness Score Prediction using Deep Learning Techniques

Ayush Nanda[0000-0002-0155-7340], Abrit Pal Singh, Aviansh Gupta,

Rajalakshmi Sivanaiah,Angel Deborah Suseelan,S Milton Rajendram,Mirnalinee T T

Department of Computer Science and Engineering
Sri SivasubramaniyaNadar College of Engineering, Chennai - 603 110, Tamil Nadu, India
{ayush18031, abritpal18007, aviansh18028}@cse.ssn.edu.in
{rajalakshmis, angeldeborahs, miltonrs, mirnalineett}@ssn.edu.in

Abstract. This paper is a description of a system used to classify tweets in Spanish as humorous or not and rate the level of humor of each tweet. The system developed by the team TECHSSN uses binary classification techniques to classify the text as humor or not (subtask1) and ensemble learning regression model to rate the funniness score of the tweet (subtask2). The data undergoes preprocessing and is given to a modification of BERT [1] (Bidirectional Encoder Representations from Transformers) for the subtask1. The model is retrained, and the weights are learned for the dataset provided. XGBoost ensemble model is used to predict the funniness score on the BERT output for subtask 2. These systems were developed for the HAHA subtasks for IberLEF2021.

Keywords: Humor Detection, Spanish, NLP, BERT.

1 Introduction

Humor is an experience that makes a person happy or amused. Throughout history, humans have been studying it from a psychological or linguistic perspective, but to see it through the eyes of a computer, which is basically figuring out the patterns and sequential repetitions in the textual content, is a challenging task for the field of NLP. One of the main reasons for this is the subjective nature of humor, as the humorousness of a joke depends on various factors such as age, gender, and cultural background of an individual. To make advancements in virtual assistants and chatbots, the integration of automated humor detection has become a necessity, which would make the conversations between them and human users more convenient and make their interactions look more human-like.We have participated in subtask 1 (humor detection) and subtask 2 (funniness score prediction).

IberLEF 2021, September 2021, Málaga, Spain.
Copyright©2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 Related Work

Humor is a well-studied topic in the fields of psychology and linguistics, but in the field of computer science continuous research has been going on and the humor-recognition systems are getting better every year for us to get a better understanding of the factors that makes a conversation humorous.

Mihalcea, R., and Strapparava, C. in their work Making Computers Laugh (2005) [2] showed that automatic classification techniques can be successfully applied to the task of humor-recognition.

UO UPV system was developed for the Humor Analysis based on Human Annotation (HAHA) track proposed in IberEval 2018 [3] Workshop. The task focuses on classifying tweets in Spanish as humorous or not and predicting how funny they are. This system combines both linguistic features and an Attention-based Recurrent Neural Network, where the attention layer helps to calculate the contribution of each term towards targeted humorous classes. This model achieves an accuracy of 84.55%.

Santiago Castro et. al. [4], in the previous iteration of our task, in IBERAMIA 2016's Natural Language Processing sub task, built a crowd sourced corpus of labeled tweets, annotated according to its humor value, letting the annotators subjectively decide which are humorous. They used SVM classifier for Spanish tweets was assembled based on supervised learning, reaching a precision of 84 % and a recall of 69 %.

In the HAHA task of IberLEF2019, Chiruzzo et. al. [5] the best classifier was developed by the user adilism [6] used the multilingual cased BERT-Base pretrained model along with the fastai library, to achieve an accuracy of 85.5% and recall of 85.2%.

Orion Weller and Kevin Seppi [7] presented a novel way of approaching this problem by building a model that learns to identify humorous jokes based on ratings learned from Reddit pages. Transformer architecture was employed using these ratings to determine the level of humor. This model outperforms all previous work done on these tasks, with an F-measure of 93.1% for the Puns dataset and 98.6% on the Short Jokes dataset.

Omar Khattab and Matei Zaharia [8] developed a novel ranking model that employs contextualized late interaction over deep language models for efficient retrieval. This architecture maintains high Mean Reciprocal Rank(MRR) at relatively lower re-ranking latency(540 times lower) and FLOPs/query(48,600 times lower) as compared to BERT-Large.

There are other papers which describe systems that detect humor in non-English text, like Ismailov A. et. al. [9], in Iberian Languages and Sushmitha Reddy Sane et. al. [10] in Hindi-English texts.

3 Methodology

For the classification of text, we have chosen the BERT-Base multilingual model which has 12 layers with the last layer's activation function as the sigmoid function, as we are performing binary classification.

3.1 Model Architecture

The classification model uses a separate line of hidden layers especially designed to extract features from each sentence. The used model is a neural network that includes two parallel lines of hidden layers: One to view text as a whole and another one to view each sentence separately. Figure 1 displays the architecture of the proposed method. It is comprised of a few general steps:

1. The sentences are separated and are tokenized individually, to analyse each sentence separately.

2. To convert the text to proper numerical inputs for the neural network, they are encoded using BERT sentence embedding. This step is performed individually on each sentence and on the whole text (shown in Figure 1).

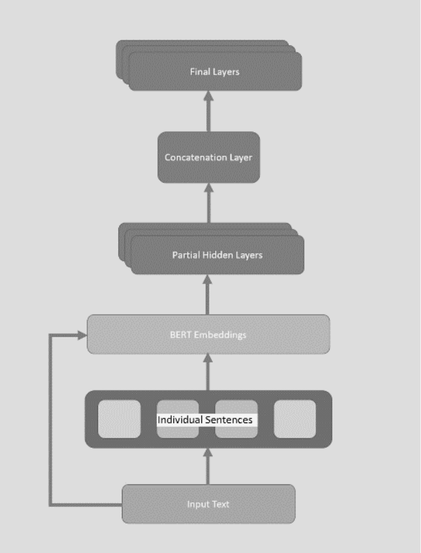

Fig. 1. Model Architecture

3. The resultant BERT sentence embeddings for each sentence that we get from the previous step are then given as an input to the partial hidden layers of the neural

network, whose purpose is to extract mid-level features for each sentence (could be related to context, type of sentence, etc).

4. While our main idea is to detect relationships between sentences (especially with punchline), it is also required to examine word-level connections in the whole text (such as synonyms and antonyms) that may have meaningful impacts in determining congruity of the text. Like the previous step, we feed BERT sentence embeddings for the whole text into hidden layers of the neural network.

5. Finally, three sequential layers of the neural network conclude our model. These final layers combine the output of all previous lines of hidden layers to result in the final output. In theory, these final layers should determine the congruity of sentences and detect the transformation of reader's viewpoint after reading the punchline.

6. For predicting the humor level (funniness score) of a tweet we factor in the votes (votes_no to votes5) instead of the binary labels, which would give us a 5-dimensional vector as a result, which is given as an input to a XGBoost Regression Model to predict the humor rating of the given tweet.

3.2 Dataset Collection

The dataset used for training our model is the one provided in the Codalab competition page, Training Dataset (haha_2021_train.csv).

Table 1. Training dataset details

Label – 'is_humor'	No. of samples
Humorous	9253
Non-Humorous	14747
Total	24000

It consists of the labels –

- **Id -** tweetId
- **text -** tweet
- **is_humor –** 0 or 1
- **votes_no, votes_1...votes_5 –** 0-1
- **humor_rating –** 1-5

humor_mechanism - {absurd, analogy, embarrassment, exaggeration, insults, irony, misunderstanding, parody, reference, stereotype, unmasking}

- **humor_target** – {age, body shaming, ethnicity/origin, family/relationships, health, lgbt, men, professions, religion, self-deprecating, sexual aggressors, social status, substance use, technology, women}

3.3 Data Preprocessing and Tokenization

For the pre-processing the data is tokenized using the BERT Tokenizer (pre-trained on the BERT-Base multilingual model) and then it undergoes stemming (SnowballStemmer) and lemmatizing (WordNetLemmatizer). The tokenized input is encoded into ids, masks, and segments, for the transformer (BERT) to accept it as an input.

3.4 Classifiers

For the model we have chosen BERT-Base multilingual, which is pre-trained in English language. This model is compared against some of the existing techniques such as Support Vector Machine (SVM), Decision Trees (DT) and Multinomial Naïve bayes (MNB).

3.5 Training, Cross Validation and Testing

For training of the model, we loop over the folds in gkf (Group K-Fold) and train each fold for 3 epochs with a learning rate of 3e-5 and a batch size of 6.As we have performed binary classification for the humor detection task, we have set the loss function as a simple binary cross-entropy, and we have chosen Adam as the optimization function. For the second task the loss function is a Mean Squared Error (MSE) function, and the optimization function remains the same.

4 Results and Analysis

As mentioned in section 3.4, the performance of our model (BERT) was tested against various machine learning techniques (SVM, DT, MNB). These models were tested on the test set with gold labels provided in the Codalab's competition page.

Table 2. Results for various models used for Subtask 1 - Humor Detection.

No.	Model	F1 Score	Accuracy	Precision	Recall
1	Our Model (BERT)	**0.7679**	**0.7978**	**0.9253**	**0.648**
2	DT	0.6121	0.6578	0.7064	0.540
3	SVM	0.5060	0.6388	0.8002	0.370
4	MNB	0.3366	0.5881	0.8648	0.209

Table 2 shows the results for the various models used for the subtask 1. This table we can infer that there is a trend, where all the models show high precision and low recall. The model has a very low false positive rate and an average false negative rate, which is illustrated in the Figure 2. And the same trend is seen amongst other models, which means that classifying a text as non-humorous is harder as compared to classifying it as humorous. Figure 2 shows the confusion matrix formed for the BERT model.

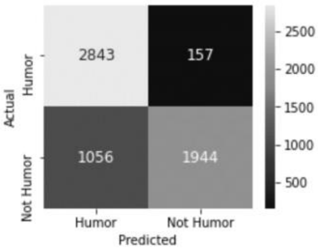

Fig. 2. Confusion Matrix of our Model
for the humor detection task

Figure 3 shows the comparison of the results for subtask 1 and 2 (humor detection and funniness score prediction) with the best approach and baseline approach of IberLEF 2021 HAHA task. We were ranked 13 in subtask 1 and 5 in subtask 2. Our model has achieved an F1-Score of 0.7679 (rank 13) in Subtask 1 and RMSE of 0.6639 (rank 5) in Subtask 2. (The overall position was based on the participant's/team's rank in sub task 1).

Table 3. Comparison of our results for humor detection and funniness score prediction task. (ST1 – Subtask1, ST2 – Subtask2)

No.	**System** (Team Name)	**Subtask1** (Subtask/Overall Position)	**Subtask2** (Subtask Position)
1	ST1 Best Approach (JOCOSO)	0.8850 (1)	0.6296 (3)
2	ST2 Best Approach (UMUTeam)	0.8544 (8)	0.6226 (1)
3	Our Approach (TECHSSN)	0.7679 (13)	0.6639 (5)
4	Baseline(-)	0.6619 (16)	0.6704 (7)

5 Conclusion

Humor detecting systems in Spanish with high accuracy can help serve to the Spanish audience on various social media platforms. It can be used to make interaction with chat-bots and virtual assistants affable. The HAHA subtask 1 and 2 for IberLEF2021 involves classifying tweets in Spanish as humorous or not and rate their humor level on a particular scale. We used a model that is built on top of BERT which is used to classify such sentences (text) into humorous or not. XGBoost regression model is used to predict the humor level or funniness score in the tweet.

References

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.

2. Mihalcea, R., and Strapparava, C.: Making Computers Laugh: Investigations in Automatic Humor Recognition. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT ’05, Association for Computational Linguistics, Vancouver, British Columbia, Canada (2005), pp. 531–538.

3. Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso: UO UPV: Deep linguistic humor detection in spanish social media. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian LanguagesI-berEval 2018 co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), pp 204–213.

4. Castro S., Cubero M., Garat D., Moncecchi G. (2016) Is This a Joke? Detecting Humor in Spanish Tweets. In: Montes y Gómez M., Escalante H., Segura A., Murillo J. (eds) Advances in Artificial Intelligence - IBERAMIA 2016. IBERAMIA 2016. Lecture Notes in Computer Science, vol 10022. Springer, Cham, pp 139-150.

5. Chiruzzo Luis., and Castro Santiago., Góngora Santiago., Rosá Aiala., Meaney, J. A. and Mihalcea Rada (2021). Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. Procesamiento del Lenguaje Natural, vol 67.

6. Orion Weller and Kevin Seppi: Humor Detection: A Transformer Gets the Last Laugh: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP):2019.

7. 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), pp 113–117.

8. Omar Khattab & Matei Zaharia. 2020.ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 39–48.

9. Ismailov, A.: Humor Analysis Based on Human Annotation Challenge at IberLEF 2019: First-place Solution. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019).

10. Sushmitha Reddy Sane, Suraj Tripathi, Koushik Reddy Sane, and Radhika Mamidi. 2019. Deep learning techniques for humor detection in hindi-english codemixed tweets. In Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp 57–61.