HAHA@IberLEF2021: Humor Analysis using Ensembles of Simple Transformers

Karish Grover[1*] and Tanishq Goel [2*]

[1] Indraprastha Institute of Information Technology, Delhi, India
karish19471@iiitd.ac.in
[2] International Institute of Information Technology, Hyderabad, India
tanishq.goel@research.iiit.ac.in

*All authors contributed equally and are mentioned alphabetically.

Abstract. This paper describes the system submitted to the Humor Analysis based on Human Annotation (HAHA) task at IberLEF 2021. This system achieves the winning F1 score of 0.8850 in the main task of binary classification (Task 1) utilizing an ensemble of a pre-trained multilingual BERT, pre-trained Spanish BERT (BETO), RoBERTa, and a naive Bayes classifier. We also achieve second place with macro F1 Scores of 0.2916 and 0.3578 in Multi-class Classification and Multi-label Classification tasks, respectively, and third place with an RMSE score of 0.6295 in the Regression task.

Keywords: Natural Language Processing · Ensemble Learning · Humor Classification · Pre-trained Models

1 Introduction

Humor Analysis based on Human Annotation (HAHA) 2021 [1] is a challenge that aims to classify Spanish tweets as humorous or not and further analyze humor by determining the characteristics present in the tweets which contribute to the humor. This challenge proposes four tasks: to classify the corpus as humorous or not, rating the humor present in the tweets, multi-class classification to find humor mechanism, and Multi-label classification tasks to find the humor target.

2 Related Work

2.1 Humor Recognition and Rating

Deep learning approaches in humor recognition have become ubiquitous like in (Chen and Soo, 2018)[2] and (Wang et al., 2020)[3]. (Weller and Seppi, 2019)[4] first proposed the use of transformers in humor detection. *Ismailov*[5] and *Annamoradnejad*[6] extended the use of BERT models to humor classification.

IberLEF 2021, September 2021, Málaga, Spain.

2.2 Voting and Ensemble Learning

Incorporating voting in ensembles is a machine learning algorithm. These algorithms have been utilized in various domains ranging from Early diabetes prediction, heart diseases prediction [7] to fields of NLP for Named Entity Recognition.

3 Data

We were provided with a corpus of crowd-annotated tweets separated into three subsets: training (24,000 tweets), development (6,000 tweets), and testing (6,000 tweets).
The columns present in the corpus utilized for training and testing are as follows:

- **text** - Text of the tweet.
- **is-humor** - binary value (0 or 1) indicating if the tweet is humorous or not.
- **humor-rating** - Real value (between 1 and 5) representing the average score the annotators gave to the tweet.
- **humor-mechanism** - Label for humor mechanism. Only a subset of the tweets have the humor mechanism annotated.
- **humor-target** - Zero or more labels for humor target, separated by ";".

4 Task Description

This challenge[3] proposes four sub-tasks which are as follows:

Humor Detection: The main aim is to classify if a tweet is humorous.

Funniness Score Prediction: Regression task which aims to rate a tweet in terms of humor.

Humor Mechanism Classification: A multi-class classification task with the primary goal of predicting the mechanism by which the tweet conveys humor.

Humor Target Classification: A multi-label classification task which aims at exploring the content of the joke based on its target.

5 Methodology

We have released our code[4] and experiments for easy replication. All the following models were fine-tuned using the **AdamW** optimizer, with a learning rate of **4e-5** and batch size of **8**. These models were trained on the NVIDIA Tesla T4 GPU.

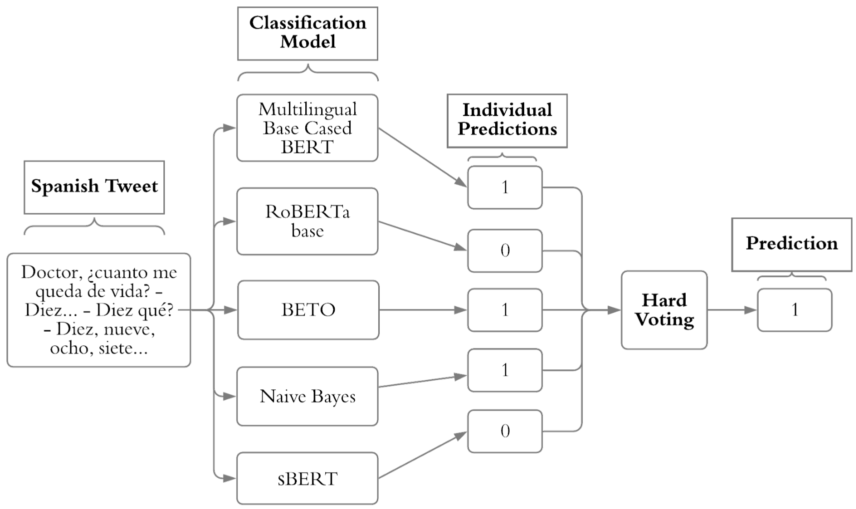

Fig. 1. Final Ensemble Model for Binary Classification Task (Task 1)

5.1 Task 1 : Binary Classification

The results for this task are summarized in Table 2. The baseline provided by the organizers for this task uses Naive Bayes with TFIDF features for Binary Classification of tweets achieving an F1 score of 0.6619 over the testing corpus.

In the final solution, we tried a series of ensembles of pre-trained models. We use the Simple Transformers classification model, **ClassificationModel** for this task which uses a pre-trained model for this task of Binary Classification.

Table 1. Ensemble Models (Experimentation)

Ensemble ID	Ensembles Used
Jocoso$_{(1)}$	sBERT + mBERT + BETO + RoBERTa + NB
Jocoso$_{(2)}$	sBERT + mBERT + ALBERT + BETO + NB + RoBERTa
Jocoso$_{(3)}$	sBERT + mBERT + BETO + NB
Jocoso$_{(4)}$	sBERT + mBERT +ALBERT + BETO + NB
Jocoso$_{(5)}$	mBERT + BETO + sBERT + DeBERTa[8] + NB
Jocoso$_{(6)}$	mBERT + BETO + ALBERT + sBERT

[3] `https://www.fing.edu.uy/inco/grupos/pln/haha/`

[4] `https://github.com/TanishqGoel/HAHA-IberLEF2021_Jocoso`

The final model is based on hard voting in an ensemble of 5 models:- Multilingual cased BERT **(mBERT)** [9] which was pre-trained on 104 languages including Spanish; **BETO** [10], which is a BERT model pre-trained on a big Spanish corpus[10]. **ALBERT** , which was pre-trained on the English language using a masked language modeling (MLM) objective; a variant of BETO model fine-tuned for sentiment analysis (**sBETO**), trained with TASS 2020 corpus (around 5000 tweets) of several dialects of Spanish. **RoBERTa** base, which is a model pre-trained on a large corpus of English data in a self-supervised fashion. Finally, we use a Multinomial **Naive Bayes Classifier** using TFIDF features. We use the Tensorflow implementation available on HuggingFace[5]. All the models were fine-tuned for **3 epochs** and took approximately 18-20 minutes for the complete training process per model.

Table 2. Task 1 (Results and Experimentation)

Ensemble	F1	Precision	Recall	Accuracy
$\text{Jocoso}_{(1)}$	**0.8850**	0.9198	**0.8526**	**0.8891**
$\text{Jocoso}_{(2)}$	0.8826	0.9194	0.8486	0.8871
$\text{Jocoso}_{(3)}$	0.8822	0.9157	0.8509	0.8863
$\text{Jocoso}_{(4)}$	0.8791	0.9176	0.8436	0.8840
$\text{Jocoso}_{(5)}$	0.8777	**0.9221**	0.8373	0.8833
$\text{Jocoso}_{(6)}$	0.8758	0.9215	0.8343	0.8816
Second Place	0.8716	-	-	-
Third Place	0.8696	-	-	-
BETO	0.8687	0.9044	0.8356	0.8736
mBERT	0.8561	0.9137	0.8053	0.8646
Baseline	0.6619	-	-	-

While training our models on the given 24,000 tweets, we observed that **BETO** outperforms all other pre-trained models. We experimented with various ensembles from these pre-trained models based on hard voting. We used a 90:10 split for the training corpus without any preprocessing. [2] We have solved this problem with the technique of classification voting ensemble, predicting the results based on the majority vote of contributing models (preference is given to BETO and multilingual BERT with high individual F1 scores).

5.2 Task 2 : Regression

The results for this task are summarized in Table 3. Here the baseline is SVM with TFIDF features which achieves an RMSE of 0.6704 over the test corpus.

[5] `https://huggingface.co/models`

[2] We observe that preprocessing reduces the F1 score.

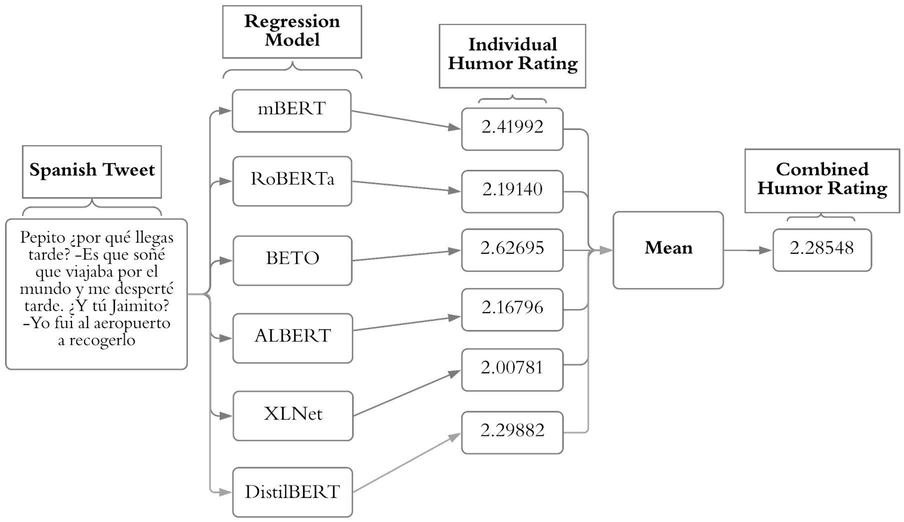

Fig. 2. Final Ensemble Model for Regression Task (Task 2)

Table 3. Task 2 (Results and Experimentation)

Ensemble	**RMSE**
First Place Solution	0.6226
Second Place Solution	0.6246
mBERT+ ALBERT + RoBERTa + DistilBERT + BETO + XLNet	**0.6295**
BETO + mBERT + ALBERT	0.6378
BETO + DistilBERT	0.6397
BETO + ALBERT	0.6391
BETO + XLNet	0.6400
BETO + mBERT	0.6412
Fourth Place Solution	0.6587
Baseline	0.6704

In this task, we tried a series of ensembles of pre-trained models, and results are predicted utilizing the technique of regression voting ensembles. We combine our model with a regression head. Our ensemble comprises of 6 pre-trained models:- Multilingual Base cased BERT (**mBERT**), **ALBERT** base v2, **RoBERTa** base, **DistilBERT** base cased [11], **BETO** [10] and **XLNet** [12] base cased model followed by regression voting. All the models were fine-tuned for **3 epochs** and took approximately 10 minutes for the complete training process per model.

5.3 Task 3 : Multi-Class Classification

The results of task 3 are summarized in table 4. The baseline provided by the organizers for Task 3 achieves a macro F1 score of 0.1001 over the training corpus, which is based on Naive Bayes with TFIDF features.

Table 4. Task 3 (Results and Experimentation)

Models Used	Macro F1 Score
First Place Solution	0.3396
BETO - Cased	**0.2916**
BETO - Cased + BETO - Uncased[6]	0.2636
Third Place Solution	0.2522
Baseline	0.1001

Our model, with a Macro F1 score of **0.2916**, utilizes BETO [10] to solve this problem of multi-class classification. We fine-tuned our model over the training corpus, which comprises of approx 4800 tweets for this task. All the models were fine-tuned for **3 epochs** and took approximately 4-5 minutes for the complete training process per model.

5.4 Task 4 : Multi-Label Classification

Table 5. Task 4 (Results and Experimentation)

Models Used	Macro F1 Score
First Place Solution	0.4228
BETO - Cased, Not Preprocessed	**0.3578**
BETO - Cased, Preprocessed[7]	0.3569
Third Place Solution	0.3225
Baseline	0.0527

Table 5 comprises the results achieved by our various ensembles and the main Spanish BERT model in task 4. The baseline mentioned lies in the range of (0.05-0.06), which is based on the frequencies of words related to a specific tag.

We use **MultiLabelClassificationModel** from Simple Transformers for this task. Our final system comprises of a pre-trained Spanish BERT cased

[6] **Combining BETO Cased and Uncased**: The BETO model classifier outputs Softmax probabilities for all the classes. We choose the top 3 classes i.e. the classes with the highest probabilities for both the models. Next, from these 6 classes, we choose the class which appears maximum times as the final prediction.

model, which is fine-tuned for 4 epochs on approximately 2000 tweets. It took approximately 5 minutes for the complete training process per model. Various ensembles and their results are listed in the above table.

6 Conclusion

This paper describes the winning solution for Task 1, the second-place solution for task 3 and task 4, and the third-place solution for Task 2 in the evaluation phase of the Humor Analysis based on Human Annotation (HAHA) challenge at the Iberian Languages Evaluation Forum (IberLEF) 2021. During the development phase, our models achieved first place in all four tasks. The combined results for both phases are mentioned in Table 4.

Table 6. Results for both Phases

Phase	Task 1	Task 2	Task 3	Task 4
Development Phase	0.8278	0.6262	0.2760	0.3389
Evaluation Phase	0.8850	0.6296	0.2916	0.3578

In all the tasks, we tried to exploit the power of voting in ensembles to get excellent results. For Task 1, 6 of our ensemble models outperform the second and third place solutions. Similarly, in other tasks, our models outperform the next place solutions by a high margin.

Further work can be done in preprocessing the Spanish tweets to analyze the effects of various preprocessing methods on Humor prediction. An interesting approach is the translation of Spanish tweets to English and back to Spanish (i.e., Back Translation) as a method of preprocessing, which is a domain open for further experimenting and research.

[7] Pre-processing includes cleaning, tokenizing, and parsing:- URLs, hashtags, mentions, reserved words (RT, FAV), emojis, and smileys. Sample preprocessor can be found at `https://pypi.org/project/tweet-preprocessor/`

References

[1] Luis Chiruzzo, Santiago Castro, Santiago Góngora, Aiala Rosá, J. A. Meaney, and Rada Mihalcea. Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. *Procesamiento del Lenguaje Natural*, 67(0), 2021.

[2] Peng-Yu Chen and Von-Wun Soo. 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, June 2018. Association for Computational Linguistics.

[3] Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. Unified humor detection based on sentence-pair augmentation and transfer learning. In *EAMT*, 2020.

[4] Orion Weller and Kevin Seppi. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Adilzhan Ismailov. Humor analysis based on human annotation challenge at iberlef 2019: First-place solution. In *IberLEF@SEPLN*, 2019.

[6] Issa Annamoradnejad and Gohar Zoghi. Colbert: Using bert sentence embedding for humor detection, 2021.

[7] Khalid Raza. Chapter 8 - improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In Nilanjan Dey, Amira S. Ashour, Simon James Fong, and Surekha Borra, editors, *U-Healthcare Monitoring Systems*, Advances in Ubiquitous Sensing Applications for Healthcare, pages 179–196. Academic Press, 2019.

[8] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[10] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

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

[12] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.