Humor Analysis based on Human Annotation(HAHA)-2019: Humor Analysis at Tweet Level using Deep Learning

Avishek Garain

Avishek Garain Computer Science and Engineering
Jadavpur University, Kolkata
avishekgarain@gmail.com

Abstract. This paper is a description of the system submitted to "Humor Analysis based on Human Annotation(HAHA)-2019" shared task. The task is divided into two sub-tasks which includes detection of humour in Spanish tweets and predicting a Humor score for the same. The tweets are short (up to 240 characters) and the language is informal, i.e., it contains spelling mistakes, emojis, emoticons, onomatopeias etc. Humor detection includes classification of the tweets into 2 classes, viz., Humorous, Not humorous. For preparing the proposed system, I use Deep Learning networks like LSTMs.

Keywords: BiLSTM · Embedding · Humor Analysis · Emoticons · Humor Score · Weighted Average

1 Introduction

Humour **D**etection refers to the use of **N**atural **L**anguage **P**rocessing (NLP) to systematically identify, extract, quantify, and study effective states and subjective information. The Humor Analysis based on Human Annotations(HAHA)-2019 was a classification task where it was required to classify a Spanish tweet on basis of its humor content,into various classes like, Humorous and Non-Humorous, and thereby predicting the Humor score if humorous. However, the task threw some additional challenges. The given tweets involved lack of context, where the number of words were less than 240. Moreover, the tweets were in an informal language and contained multi-linguality. Also, the classification system that would be prepared for the task, needed to be generalized for various test corpora as well.

To solve the task in hand, I built a bidirectional **L**ong **S**hort **T**erm **M**emory (LSTM) based neural network, for classification purpose as well as humor score prediction purpose.

The rest of the paper has been organized as follows. Section 2 describes the data, on which, the task was performed. The methodology followed is described in Section 3. This is followed by the results and concluding remarks in Section 4 and 5 respectively.

2 Data

The dataset that was used to train and validate the model was provided by the IberLEF [2]. The data was collected from Twitter and it was retrieved using the Twitter API by searching for keywords and constructions that are often included in various texts of different sentiments. The dataset provided consisted of tweets in their original form along with the corresponding labels and scores, as shown in Table 1.

Label value	Meaning
0	Not Humorous
1	Humorous

Table 1: Labels used in the dataset

The dataset originally comprised of Spanish tweets. The tweets were also tagged with their respective Humor labels. The resulting dataset had 24,000 humor tagged and scored tweets, which were splitted into 16,800 instances of training data and 7,200 instances of development data. My approach was to convert the tweets into a sequence of words and convert them into word embeddings. I then ran a neural-network based algorithm on the processed tweet. Language and label based categorical division of data is given in Table 2, 3 and 4.

Value	Humorous	Not-Humorous
All	10323	6477

Table 2: Distribution of the labels in the training dataset

Value	Humorous	Not-Humorous
All	4424	2776

Table 3: Distribution of the labels in the development dataset

Value	Humorous	Not-Humorous
All	14747	9253

Table 4: Distribution of the labels in the combined dataset

3 Methodology

I used SenticNet5[1] for finding sentiment values of individual words after converting the sentences to English via GoogleTrans API. Apart from this, I also used a Spanish Sentiment lexicon for the same.

The use of BiLSTM networks is a key factor in our model. The work of [5] brought a revolutionary change by bringing the concept of memory into usage for

sequence based problems.
I first took the tweets and sent the raw data through some preprocessing steps, for which I took inspiration from the work on Hate Speech against immigrants in Twitter[4], part of SemEval2019. The steps used here are built as an advancement of this work. It consisted of the following steps:

1. Replacing emojis and emoticons by their corresponding meanings
2. Removing mentions
3. Removing URLs
4. Contracting whitespace
5. Extracting words from hashtags

In step 1, for example,
”,-) ” is replaced by ”winking happy”
”;-(” is replaced with ”crying”
”:-C” is replaced with ”real unhappy”
Similarly I replaced 110 emoticons by their feelings.

The last step (step 5) consists of taking advantage of the Pascal Casing of hashtags (e.g. `#AngryBird`). A simple regular expression could extract all words; I ignored a few errors that arise in this procedure. Using this extraction, contributed to features mainly because words in hashtags, to some extent, may convey sentiments for the tweet. They played an important role during the model-training stage.

The preprocessed tweets are treated as a sequence of words with interdependence among various words contributing to its meaning. I convert the tweets into one-hot vectors. I also included certain manually extracted features listed below:

1. Counts of words with positive sentiment, negative sentiment and neutral sentiment in Spanish
2. Counts of words with positive sentiment, negative sentiment and neutral sentiment in English
3. Subjectivity score of the tweet
4. Number of question marks,Exclamations and full-stops in the tweet

I use a Bidirectional-LSTM based approach to capture information from both the past and future context.

My models for both the sub-tasks are neural-network based models. For subtask-1, first, I merged the manually-extracted features with the feature vector obtained after converting the processed tweet to one-hot encoding. The output was processed through an embedding layer which transformed the tweet into a 128 length vector. The embedding layer learns the word embeddings from the input tweets. I passed the embeddings through a Bidirectional LSTM layer containing 128 units. This was followed by another bidirectional LSTM layer containing 256 units with its dropout and regular dropout set to 0.45 and activation being a sigmoid activation. This is followed by a Bidirectional LSTM layer with 128 units for better learning. This was followed by the final output layer of neurons with sigmoid activation, where, each neuron predicts a label as present in the dataset.

For sub-task 1, I trained a model containing 2 neurons for predicting `Humorous`, `Not humorous` respectively. The model was compiled using the Adam optimization

algorithm with a learning rate of 0.0005. Binary-crossentropy was used as the loss function. The working is depicted in Figure 1.

For sub-task 2, I trained on the same model but the Dense layer consisted of 5 neurons for five classes representing the classes ranging from 1-star count to 5-star count. I got a probability prediction percentage against each of the classes and thus finally got the final humour score by finding the Weighted Average using the following formula:

$$S = \frac{\sum_{i=1}^{5}(p_i * i)}{\sum_{i=1}^{5} p_i}$$

where,
p_i=Probability of getting i stars humor rating
i=Number of stars
The model is compiled using the Adam optimization algorithm with a learning rate of 0.001. Categorical crossentropy is used as the loss function.

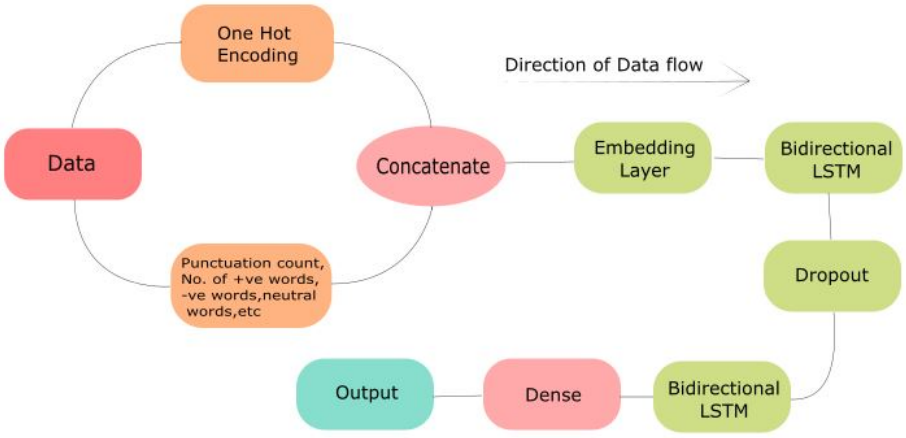

Fig. 1: Flowchart of working model

I noted that the dataset is highly skewed in nature. If trained on the entire training dataset without any validation, the model tended to completely overfit to the class with higher frequency as it led to a higher accuracy score.

To overcome this problem, I took some measures. Firstly, the training data was splitted into two parts — one for training and one for validation comprising 70 % and 30 % of the dataset respectively. The training was stopped when two consecutive epochs increased the measured loss function value and decrease in Validation accuracy for the validation set.

Secondly, class weights were assigned to the different classes present in the data which were chosen to be proportional to the inverse of the respective frequencies of the classes. Hypothetically, the model then gave equal weight to the skewed classes and this penalized tendencies to overfit to the data.

4 Results

I participated in subtasks 1 and 2 of Humor Analysis based on Human Annotation(HAHA)-2019 and our system works quite well.

I have included the automatically generated tables of evaluation metrics with my results. The results are depicted in Tables 5-7.

System	Train (%)	Validation (%)
Without	85.13	76.84
With	90.32	80.64

Table 5: Comparison of development phase accuracies with and without hashtag preprocessing

Task-1

System	F1	Precision	Recall
BiLSTM	0.593	0.491	0.748

Table 6: Result Metrics

Task-2

System	RMSE
BiLSTM	1.653

Table 7: Root Mean Square Error

5 Conclusion

In this system report, I have presented a model which performs satisfactorily in the given tasks. The model is based on a simple architecture. There is scope for improvement by including more manually extracted features (like those removed in the preprocessing step) to increase the performance. Another fact is that the model is a constrained system, which may lead to poor results based on the modest size of the data. Related domain knowledge may be exploited to obtain better results. Use of regularizers led to proper generalization of model, henceforth increasing our task submission score.

References

1. Cambria, E., Poria, S., Hazarika, D., Kwok, K.: Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In: AAAI (2018)
2. Castro, S., Chiruzzo, L., Rosá, A., Garat, D., Moncecchi, G.: A crowd-annotated spanish corpus for humor analysis. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media. pp. 7–11 (2018)
3. Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J.J., Rosá, A.: Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
4. Garain, A., Basu, A.: The titans at SemEval-2019 task 5: Detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 494–497. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019), https://www.aclweb.org/anthology/S19-2088
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, http://dx.doi.org/10.1162/neco.1997.9.8.1735
6. Mihalcea, R., 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. pp. 531–538. HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005). https://doi.org/10.3115/1220575.1220642, https://doi.org/10.3115/1220575.1220642
7. Sjöbergh, J., Araki, K.: Recognizing humor without recognizing meaning. In: Masulli, F., Mitra, S., Pasi, G. (eds.) Applications of Fuzzy Sets Theory. pp. 469–476. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)