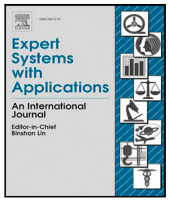
Expert Systems With Applications 249 (2024) 123685

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



# ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor

Issa Annamoradnejad [a,*], Gohar Zoghi [b]

[a] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[b] Golestan University of Medical Sciences, Gorgan, Iran

ARTICLE INFO

Dataset link: https://github.com/Moradnejad/ColBERT-Using-BERT-Sentence-Embedding-for-Humor-Detection

*Keywords:*
Computational humor
Humor detection
Humor rating
Parallel neural networks
BERT sentence embedding
Jokes dataset

ABSTRACT

Automatic humor detection has compelling use cases in modern technologies, such as humanoid robots, chatbots, and virtual assistants. In this paper, we propose a novel approach for detecting and rating humor in short texts based on a popular linguistic theory of humor. The proposed technical method initiates by separating sentences of the given text and utilizing the BERT model to generate embeddings for each one. The embeddings are fed to a neural network as parallel lines of hidden layers in order to determine the congruity and other latent relationships between the sentences, and eventually, predict humor in the text. We accompany the paper with a novel dataset consisting of 200,000 short texts, labeled for the binary task of humor detection. In addition to evaluating our work on the novel dataset, we participated in a live machine-learning competition to rate humor in Spanish tweets. The proposed model obtained F1 scores of 0.982 and 0.869 in the performed experiments which outperform general and state-of-the-art models. The evaluation results confirm the model's strength and robustness and suggest two important factors in achieving high accuracy in the current task: (1) usage of sentence embeddings and (2) utilizing the linguistic structure of humor in designing the proposed model.

## 1. Introduction

For ages, human beings have been fantasizing about humanoid robots indistinguishable from humans. In making that a reality, humor cannot be missed as a major human feature, which for its subjectivity, ambiguity, and semantic intricacies has been a difficult problem for researchers to tackle. As an example, In Interstellar (2014 movie), a future earth is depicted where robots easily understand and use humor in their connections with their owners and humans can set the level of humor in their personal robots.[1] While we may have a long path toward astral travels, we are very close to reaching high-quality artificial intelligence systems (such as chatbots, virtual assistants, and even robots) injected with adjustable humor. This work contributes to this human fantasy and paves the way for creating such systems.

Humor, as a potential cause of laughter, is an important part of human communication, which not only makes people feel comfortable but also creates a cozier environment (Castro, Cubero, Garat, & Moncecchi, 2016). Automatic humor detection has interesting use cases in building human-centered artificial intelligence systems such as humanoid robots, chatbots, and virtual assistants. An appealing use case is to identify whether an input command should be taken seriously or not, which is a critical step to understanding the real motives of users, returning appropriate answers, and enhancing the overall experience of users with the AI system. A more advanced outcome would be the injection of humor into computer-generated responses, thus making the human–computer interaction more engaging and interesting (Niculescu, van Dijk, Nijholt, Li, & See, 2013). This is an outcome that is achievable by setting the level of humor in possible answers to the desired level, similar to the mentioned movie.

Humor can be attained through several linguistic or semantic mechanisms, such as wordplay, exaggeration, misunderstanding, and stereotyping. Researchers proposed several theories to explain humor functionality as a trait, one of which is called "incongruity theory", in which laughter is the result of realizing incongruity in the narrative. A general version states that a common joke consists of a few sentences that conclude with a punchline. The punchline is responsible for bringing contradiction into the story, thus making the whole text laughable. In other words, any sentence can be non-humorous in itself, but when one tries to comprehend all sentences together in one context or as a whole, the text becomes humorous.

The tasks of humor classification and generation have gained increasing attention in recent years (Meaney, 2020; Peyrard, Borges,

* Corresponding author.
*E-mail addresses:* i.moradnejad@gmail.com (I. Annamoradnejad), zoghi.g@goums.ac.ir (G. Zoghi).

[1] Tarzs, in the movie.

https://doi.org/10.1016/j.eswa.2024.123685
Received 10 March 2023; Received in revised form 13 January 2024; Accepted 13 March 2024
Available online 19 March 2024

Gligorić, & West, 2021; Weller & Seppi, 2019; Ziser, Kravi, & Carmel, 2020). These tasks have been the subject of research for some time, with early attempts by Mihalcea and Strapparava (2005), who proposed combining humor-specific stylistic features and content-based features to classify short sentences. Purandare and Litman (2006) explored the use of acoustic-prosodic features, such as pitch and energy, in addition to linguistic features in spoken conversations.

In this work, however, the problem is tackled by considering the underlying structure of humor in texts, in addition to utilizing recent language models for generating strong sentence embeddings. By focusing on the mentioned linguistic structure of jokes, we suggest and show that it is required to view and encode each sentence separately and capture the underlying relation between sentences in a proper way. As a result, our proposed model for the task of humor detection is based on creating parallel paths of neural network hidden layers, in addition to encoding a given text as a whole.

The proposed approach initiates by separating the text into its sentences. It continues by utilizing the BERT model to encode each sentence and the whole text as embeddings. Next, the embeddings will be fed into parallel hidden layers of a neural network to extract latent features regarding each sentence. The last three layers combine the output of all previous lines of hidden layers to determine the relationship between the sentences to predict the final output. In theory, these final layers are responsible for determining the congruity or detecting the transformation of the reader's viewpoint after reading the punchline.

To test the robustness and stability of the proposed model, its performance is evaluated in two different settings:

1. Evaluation on a new dataset (short informational English texts): The model is evaluated for the binary task of humor detection on the novel dataset, and
2. Evaluation in a live machine-learning competition (Spanish informal texts): The model competes against strong real teams to detect and rate humor in informal Spanish tweets (variant lengths).

One of the main problems of automatic humor detection is the subjectivity of humor. What one person finds funny may not be the same for another person, and it can be difficult for a machine to accurately determine what is humorous to a particular individual. This subjectivity can make it difficult to accurately evaluate the performance of a humor detection model, as it is not clear what the "correct" answer should be in many cases. Thus, it can be challenging for a machine to accurately capture these varied definitions. This lack of a clear definition can also make it difficult to evaluate the performance of a humor detection model, as it is not always clear what the model is trying to detect. In addition, previous attempts to create a dataset for humor detection mostly combined formal and informational non-humorous texts with informal conversational humorous short texts, which due to the incompatible statistics of the parts (text length, words count, etc.), makes it more likely to detect humor with simple analytical models and without understanding the underlying latent lingual features and structures.

To address these problems, we curated a large dataset for the binary task of humor detection. The new dataset may provide a more reliable resource for researchers working on humor detection and may facilitate further progress in the field. Even by focusing to fix these issues, as discussed in Section 5.1.2, the novel dataset or any binary classification dataset would contain a few examples in the gray area due to the subjectivity of humor. In the second evaluation, this problem is mitigated as the task is to rate the level of humor (a regression task) on a dataset that is labeled by multiple anonymous annotators.

This work advances the state of the art in humor detection and represents a significant contribution to the field. We summarize our contributions as follows:

- The ColBERT[2] model is proposed for the task of humor detection and rating based on a general linguistic theory of humor. The model architecture and components are presented in detail.
- A large novel dataset, entitled the "ColBERT dataset", is curated, labeled, and published for the task of humor detection. The dataset contains 200k short texts (100k positive and 100k negative). The focus was to reduce or completely remove issues prevalent in the existing datasets of the binary task.
- The performance of the proposed method is evaluated on the novel dataset in comparison with five strong baselines.
- We further evaluate its accuracy and robustness in a data science competition on humor detection and rating of Spanish texts.

The structure of this article is as follows: Section 2 reviews linguistic structure of humor and recent works on the task of humor detection with a focus on transfer learning methods. Section 3 describes the data collection and preparation process utilized for the novel dataset. Section 4 elaborates on the methodology, and Section 5 presents our experimental results. Section 6 is the concluding remarks.

## 2. Background

Humor has been a topic of research in various fields of science, including psychology, linguistics, and natural language processing (NLP).

In this section, first, we discuss the general linguistic structure of a joke. On the second part, the field of computational humor with a focus on NLP and transfer learning is reviewed.

### 2.1. General linguistic structure of humor

The proposed method in this article is based on a general linguistic structure of a joke. Therefore, these theories are introduced in this part.

There has been a long line of works in linguistics of humor that classify jokes into various categories based on their structure or content. Early research work on humor can be traced back to Kline (1907) and Wolff, Smith, and Murray (1934). Along the way, many suggested that humor arises from the sudden transformation of an expectation into nothing (Kant, 1913), known as Incongruity Theory. In this way, the punchline, as the last part of a joke, destroys the perceiver's previous expectations and brings humor to its incongruity. Some have suggested that the structure of a joke involves two or three stages of storytelling that conclude with a punchline (Eysenck, 1942; Suls, 1972).

Raskin (2012) presented Semantic Script Theory of Humor (SSTH), a detailed formal semantic theory of humor. The SSTH has the necessary condition that a text has to have two distinct related scripts that are opposite in nature, such as real/unreal, and possible/impossible. For example, review a typical joke:

> "Is the doctor at home?" the patient asked in his bronchial whisper. "No", the doctor's young and pretty wife whispered in reply. "Come right in". (Raskin, 2012)

This is compatible with the two-staged theory which ends with a punchline. The punchline is related to previous sentences but is included in opposition to previous lines in order to transform the reader's expectation of the context.

It is worth mentioning that in addition to Incongruity Theory of humor that the proposed approach of this work is based on, there are other theories that discuss the cause of laughter in humans (see Attardo (2010), Scheel (2017)).

[2] Denoting the use of BERT language model in columns of neural network.

### 2.2. Computational humor

With advances in NLP, researchers applied and evaluated state-of-the-art methods for the task of humor detection. This includes using statistical and N-gram analysis (Taylor & Mazlack, 2004), Regression Trees (Purandare & Litman, 2006), Word2Vec combined with K-NN Human Centric Features (Yang, Lavie, Dyer, & Hovy, 2015), and Convolutional Neural Networks (Chen & Soo, 2018; Weller & Seppi, 2019).

Zhang and Liu (2014) tackled the problem of humor recognition in tweets using phonetic, morpho-syntactic, lexicosemantic, pragmatic, and affective features. Bertero and Fung (2016) combined hierarchical continuous representations with high-level features, such as structural features, antonyms, and sentiment, to predict the humor of body punchlines in TV sitcom dialogues. Chen and Soo (2018) proposed a CNN-based architecture combined with highway networks for humor recognition.

There are emerging tasks related to humor detection. Yang, Hu, and Hirschberg (2019) focused on predicting humor by using audio information, hence reaching 0.750 AUC[3] by using only audio data. A good number of research is focused on detecting humor in non-English texts, such as in Spanish (Chiruzzo et al., 2019; Giudice, 2019; Ismailov, 2019), Chinese (Yang, Hu, & Hirschberg, 2019), and English–Hindi (Khandelwal, Swami, Akhtar, & Shrivastava, 2018). Furthermore, some recent studies focused on detecting humor in voice, videos, and pictures (e.g. Choube and Soleymani (2020), Hasan et al. (2019), Patro et al. (2021), Wu, Lin, Yang, and Xu (2021))

With the popularity of transfer learning, some researchers focused on using pre-trained models for several tasks of text classification. BERT (Devlin, Chang, Lee, & Toutanova, 2018), a popular strong language model, utilizes a multi-layer bidirectional transformer encoder consisting of several encoders stacked together, which can learn deep bi-directional representations. Similar to previous transfer learning methods, it is pre-trained on unlabeled data to be later fine-tuned for a variety of tasks. It initially came with two model sizes ($\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$) and obtained eleven new state-of-the-art results. Since then, it was pre-trained and fine-tuned for several tasks and languages, and several BERT-based architectures and model sizes have been introduced (such as Multilingual BERT, RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) and VideoBERT (Sun, Myers, Vondrick, Murphy, & Schmid, 2019)).

Frenda et al. (2022) investigated the peculiarities of sarcasm as a type of irony by examining the IronITA dataset of Italian tweets about controversial social issues. To better understand the impact of hurtful and affective language on the detection of irony and sarcasm, the researchers developed a transformer-based system called AlBERToIS, which combined the pre-trained AlBERTo model with linguistic features. This approach performed the best on irony and sarcasm detection on the IronITA dataset.

Weller and Seppi (2019) focused on the task of humor detection by using a Transformer architecture. The work approached the task by learning on ratings taken from the popular Reddit r/Jokes thread (13884 negative and 2025 positives) and designed a BERT architecture for this task, which was able to compete with human perception. Wang, Yang, Qin, Sun, and Deng (2020) designed a multilingual model based on pre-trained BERT models for Chinese, Russian, and Spanish, which was fine-tuned on inter sentence relationship and sentence discrepancy prediction for body punchlines. Low, Fung, Iqbal, and Huang (2022) applied transformers, a recent development in neural network architecture, to the task of separating satirical news from fake news. Using a classifier framework built around a DistilBERT architecture, they achieved an improvement in F1 and accuracy.

[3] Area Under the receiver operating characteristic (ROC) Curve.

**Table 1**
Datasets for the binary task of humor classification.

| Dataset | # Positive samples | # Negative samples |
|---|---|---|
| 16000 One-Liners (Mihalcea & Strapparava, 2005) | 16,000 | 16,002 |
| Pun of the Day (Yang et al., 2015) | 2,423 | 2,403 |
| PTT Jokes (Chen & Soo, 2018) | 1,425 | 2,551 |
| English-Hindi (Khandelwal et al., 2018) | 1,755 | 1,698 |
| ColBERT | 100,000 | 100,000 |

## 3. Dataset

Existing humor detection datasets use a combination of formal and informational non-humorous texts with informal conversational jokes. This results in incompatible lexical statistics (text length, word count, etc.), making it more likely to detect humor with simple analytical models and without understanding the underlying latent connections. Moreover, the previous datasets are relatively small for the tasks of text classification, making them prone to over-fit models and result in incorrect validations. These problems encouraged us to create a new dataset exclusively for the task of humor detection; thus, simple lexical-based models will not be able to fully successful in predicting humor.

The remainder of this section elaborates on the data collection process, data sources, preprocessing methods, and statistics of the new dataset.

### 3.1. Data collection

The existing relevant datasets (exclusively on news stories, news headlines, Wikipedia pages, tweets, proverbs, and jokes) were analyzed with regard to table size, character length, word count, and formality of language. Table 1 compares of the existing humor detection datasets (binary task) with regard to their size. There are other datasets focused on closely related tasks, e.g. the tasks of punchline detection and success (whether or not a punchline triggers laughter) (Chen & Lee, 2017; Hasan et al., 2019), or on using speak audio and video to detect humor (Bertero & Fung, 2016; Hasan et al., 2019).

To create the dataset, we chose the following two data sources, both of which are formed of informational un-abbreviated texts (one with humor texts and one without):

1. News category dataset (Misra, 2018), published under CC0 Public Domain, consisting of 200k Huffington Post news headlines from 2012–2018, along with corresponding URLs, categories, and full stories. The stories are scattered in several news categories, including politics, wellness, entertainment, and parenting.
2. Jokes dataset contains 231,657 jokes/humor short texts, crawled from Reddit communities.[4] The dataset is compiled as a single `csv` file with no additional information about each text (such as the source, date, etc.) and is available at Kaggle. Chen and Soo (2018) combined this dataset with the WMT162 English news crawl but did not publicly publish the dataset. Weller and Seppi (2019) also combined this dataset with extracted sentences from the WMT162 news crawl and made it publicly available.

The items from the Jokes dataset will be considered as positive target class and the items from the news category as the negative target class.

[4] Mostly from /r/jokes and /r/cleanjokes subreddits.

**Table 2**
Textual statistics of the ColBERT dataset (100k positive, 100k negative).

| | #chars | #words | #punctuation | #duplicate words | Sentiment polarity | Sentiment subjectivity |
|---|---|---|---|---|---|---|
| mean | 71.561 | 12.811 | 2.378 | 0.440 | 0.051 | 0.317 |
| std | 12.305 | 2.307 | 1.941 | 0.794 | 0.288 | 0.327 |
| min | 36 | 10 | 0 | 0 | −1.000 | 0.000 |
| median | 71 | 12 | 2 | 0 | 0.000 | 0.268 |
| max | 99 | 22 | 37 | 13 | 1.000 | 1.000 |
| corr with target | 0.03 | 0.05 | 0.00 | 0.03 | 0.09 | 0.02 |

**Table 3**
A few examples from the novel dataset.

| Text | Is humor? |
|---|---|
| Why your purse is giving you back pain... and 11 ways to fix it | False |
| Why was the fruit/vegetable hybrid upset? he was a melon-cauliflower. | True |
| On set with Paul Mitchell: from our network | False |
| Starting a cover band called a book so no one can judge us. | True |

### 3.2. Preprocessing and filtering

In this step, a few preprocessing measures were implemented to enhance the lexical similarity of the target classes.

The initial step was to drop duplicate texts, as there were duplicate samples in both data sources. Dropping duplicate samples removed 1369 items from the jokes dataset and 1558 items from the news dataset.

To make the lexical statistics similar, the average and standard deviation of the number of characters and words for each group is calculated. Then, we selected texts in pairs in a way that each text from one dataset will have a text with the same statistics from the other class. As a result, only texts with character lengths between 30 and 100 and word lengths between 10 and 18 were remained.

In addition, we noticed that headlines in the news dataset use Title Case[5] formatting, which was not the case with the jokes dataset. Thus, Sentence Case[6] formatting was applied to all news headlines by keeping the first character of the sentences in capital and lower-casing the rest. This simple modification prevents simple classifiers from reaching perfect accuracy.

Finally, 100k samples from each data source were randomly selected and merged to create an evenly distributed dataset.

### 3.3. ColBERT dataset

The dataset[7] contains 200k labeled short texts equally distributed between humor and non-humor. It is much larger than the previous datasets (Table 1) and it includes texts with similar textual features. Table 3 contains a few random examples from the dataset.

As mentioned in 3.1, the objective of this section was to construct a dataset with two classes, ensuring that items from both classes exhibit similar lexical statistics. This objective is corroborated by the results of correlation analysis between the lexical statistics of both classes and the target column, as illustrated in Table 2. The correlation between character count and the target is marginal. Furthermore, the correlation analysis for sentiment polarity and subjectivity[8] showed −0.09 and +0.02, respectively, which suggests limited connection between the target value and sentiment features.

To facilitate future researchers in properly testing their chosen model, 40,000 samples were randomly selected and separated from the remaining dataset to constitute the test set. It is essential to note that, to prevent data leakage, the test set must remain distinct from both the training and validation sets during the configuration of hyperparameters. If necessary, the training set can be further divided into separate training and evaluation subsets.

## 4. Proposed method

This section elaborates on the proposed method.

### 4.1. Model architecture

Based on the presented linguistic structure of humor (Section 2.1), if one reads each sentence of a joke separately, it is most likely to be found as a normal and non-humorous text. On the other hand, when we try to comprehend all sentences together in one context or as a whole, the text becomes humorous. Our proposed method utilizes this linguistic characteristic of humor in order to view or encode sentences separately and extract mid-level features using hidden layers.

Fig. 1 displays the architecture of the proposed method. It contains separate paths of hidden layers specially designed to extract latent features from each sentence. Furthermore, there is a separate path to extract latent features of the whole text. Hence, our proposed neural network structure includes a single path to view the text as a whole and several other paths to view each sentence separately. It is comprised of a few general steps:

1. First, to assess each sentence separately and extract numerical features, sentences are separated and tokenized individually.
2. To prepare these textual parts as proper numerical inputs for the neural network, they are encoded using BERT sentence embedding. This step is performed individually on each sentence (left side in Fig. 1) and also on the whole text (right side in Fig. 1).
3. Now that we have BERT sentence embedding for each sentence, they are fed into parallel hidden layers of a neural network to extract mid-level features for each sentence (related to context, type of sentence, etc.). The output of this part for each sentence is a vector of size 20.
4. While our main idea is to detect existing relationships between sentences (specifically the punchline's relationship with the rest), it is also required to examine word-level connections in the whole text that may have meaningful impacts in determining the congruity of the text. For example, the existence of synonyms and antonyms in different sentences of the text could be meaningful. The BERT sentence embedding for the whole text is fed into a different line of layers (right side in Fig. 1). The output of this part is a vector of size 60.
5. Finally, three sequential layers of the neural network conclude our model. These final layers combine the output of all previous paths of hidden layers in order to predict the final output. In theory, these final layers should determine the congruity of sentences and detect the transformation of the reader's viewpoint after reading the punchline.

[5] All words are capitalized, except non-initial articles like "a, the, and", etc.
[6] Capitalization as in a standard English sentence, e.g., "Witchcraft is real.".
[7] The dataset is available at: https://github.com/Moradnejad/ColBERT-Using-BERT-Sentence-Embedding-for-Humor-Detection.
[8] Calculated for all texts using TextBlob python library.

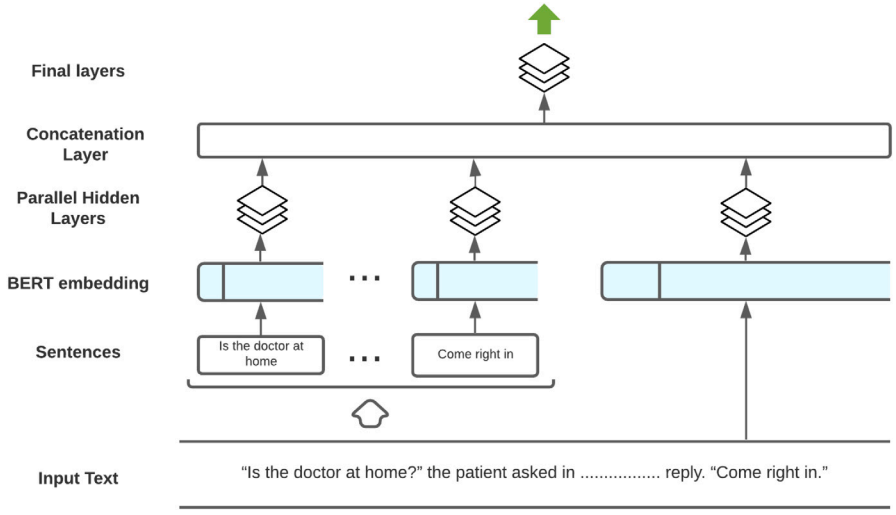



**Fig. 1.** Components of the proposed method.

### 4.2. Implementation notes

Since our approach builds on using BERT sentence embedding in a neural network, the token representation is obtained using a BERT tokenizer with a maximum sequence length of 100 (the maximum sequence length of BERT is 512). Then, the BERT sentence embedding is generated by feeding tokens as input into the BERT model (vector size=768).

The model will pass BERT embedding vectors of the given text and its sentences as inputs to a neural network with eight layers. For each sentence, there is a separate line composed of three hidden layers, parallel to other sentence lines. These lines are eventually concatenated in the fourth layer and continue in a sequential manner to predict the single target value. The `huggingface` and `keras.tensorflow` packages are used for the BERT model and neural network implementations, respectively.

It is important to note that the BERT model is used to generate sentence embedding. Therefore, training is performed on the neural network and not on the BERT model. BERT comes with two pre-trained general types (the $\text{BERT}_{\text{BASE}}$ and the $\text{BERT}_{\text{LARGE}}$), both of which are pre-trained from unlabeled data extracted from BooksCorpus (Zhu et al., 2015) with 800M words and English Wikipedia with 2,500M words (Devlin et al., 2018). We used the smaller sized model ($\text{BERT}_{\text{BASE}}$ with 12 layers, 768-hidden states, 12-heads, and 110M parameters) pre-trained on lower-cased English text (uncased).

To achieve clean data, a few textual preprocessing steps are performed on all input texts. It should be noted that these steps are performed as part of the methodology; thus, the novel dataset is not impacted. The preprocessing steps are:

- **Expanding contractions**: all contractions are replaced with the expanded version of the expressions. For example, "is not" instead of "isn't".
- **Cleaning punctuation marks**: the punctuation marks[9] are separated from words to achieve cleaner sentences. For example, the sentence "This is' (fun)". is converted to "This is ' (fun) ".
- **Cleaning special characters**: some special characters are replaced with an alias. For example, "alpha" instead of "$\alpha$".

[9] The punctuation marks are: period, comma, question mark, hyphen, dash, parentheses, apostrophe, ellipsis, quotation mark, colon, semicolon, exclamation point.

## 5. Evaluation and discussion

In this section, the performance and robustness of the proposed method is evaluated on two settings. First, the performance of the proposed model is compared against five strong baselines on the novel ColBERT dataset of formal English texts. Then, we show the robustness and good performance of the model by participating in a shared-task competition focused on detecting and rating humor in informal Spanish tweets.

### 5.1. Evaluation on the ColBERT dataset

In this part, the performance of the ColBERT model is compared with five baselines on the novel dataset.

As mentioned in Section 3.3, the data is randomly split into 80% (160K) train and 20% (40K) test parts. To prevent data leakage, the test part is not used in any of the steps regarding training, modeling, hyperparameter selection, fine-tuning, or validation. The train part is further split according to the cross-validation scheme (K=5) to find the proper hyper-parameters of the proposed and baseline models. Thus, in every fold, 128 K is used for training and the remaining 32 K for validation. Finally, a single model is trained on all samples of the train set (160K) based on the selected hyper-parameters. The reported results of this section are based on the final evaluation of the trained models on the test part of the dataset (remaining 40 K).

The baseline models are:

1. Decision Tree: A methodology that is commonly used as a data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. The method uses the train dataset to generate branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes (Song & Ying, 2015). `CountVectorizer` is used to generate numerical word representations.
2. SVM: A supervised model that achieved robust results for many classification and regression tasks. For this baseline, `TfidfVectorizer` is applied to generate numerical word representations with some optimization on hyperparameters.
3. Multinomial naïve Bayes: The model is suited when we deal with discrete integer features, such as word counts in a text. Similar to Decision Tree, `CountVectorizer` is used to generate numerical word representations.

**Table 4**
Performance evaluation on the ColBERT Dataset.

| 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 |
| ColBERT | | 0.982 | 0.990 | 0.974 | 0.982 |

4. XGBoost: XGBoost is the latest step in the evolution of tree-based algorithms that include decision trees, boosting, random forests, boosting and gradient boosting. It is an optimized distributed gradient boosting that provides fast and accurate results, which achieves accurate results in less time (Chen & Guestrin, 2016). We applied the model on numerical word representations generated by `CountVectorizer` which resulted in better accuracy than `TfidfVectorizer`.
5. XLNet: A generalized language model that aims to mitigate the issues related to the BERT model and previous autoregressive language models. For the task of text classification (and some other NLP tasks), XLNet outperforms BERT on several benchmark datasets (Yang, Dai, et al., 2019). `xlnet-large-cased` (24 layers and 340M parameters) is used for the purposes of this evaluation.

#### 5.1.1. Results

The results of our experiments on the ColBERT dataset are displayed in Table 4. The evaluations found the proposed model's accuracy and F1 score to be 98.2%, thus outperforming all selected baselines by a large margin. This is a 7% jump from the recent state-of-the-art XLNet model (with 340M parameters) and 17% higher than the gradient boosting classifier. Traditional models of Decision Tree, SVM, and Multinomial naïve Bayes gained less than 90% accuracy, still an acceptable performance for a general model. XGBoost, a strong implementation of gradient boosting achieved 81% F1-score based on the selected word representations. $\text{XLNet}_{\text{Large}}$, which required less optimization, was the strongest among the baselines, reaching close to 92% accuracy, 4 percent higher than Multinomial NB.

Regarding time performance, the proposed model requires 2 h (on average) to perform one epoch of training on 128k samples of the dataset on a computer with NVIDIA TESLA P100 GPUs. This is comparably less than the XLNet model (3.5 h), but longer than the rest of the selected baselines (0.2-2 h).

#### 5.1.2. Discussions on the first evaluation

To comprehensively assess the effectiveness of the proposed model, the Friedman rank sum test was employed within a two-way balanced complete block design. This statistical analysis aims to test significant differences in performance among our model and five baseline models. The Friedman chi-squared statistic yielded a value of 19.142857, calculated with degrees of freedom of five, representing the number of correlated samples or groups minus one. As the calculated $p$-value of 0.001808 corresponding to the omnibus null hypothesis is lower than the conventional threshold of 0.05, we have grounds to infer that one or more of the correlated samples exhibit meaningful differences.

In light of this evidence, further investigation was carried out through post-hoc pairwise multiple comparison tests utilizing the Conover method. The derived p-values underwent adjustments using the Benjamini-Hochberg false discovery rate (FDR) procedure, providing a refined perspective on model performance. The tests resulted in Conover p-values of 0.070, 0.220, 0.291, 0.069, and 0.570 for comparisons between our proposed method and the five baseline models[10].

[10] With the same order of models in Table 4.

These results elucidate notable statistical differences in performance, especially with the first four baselines. The $p$-value of 0.570 implies a comparably higher similarity in performance between our method and XLNet, as seen in the metrics.

The results suggest two important factors in achieving high accuracy in the current task. First, methods that rely on pre-trained language models to produce sentence embedding outperform traditional methods of the word representation. The XLNet and proposed models both use their own embeddings and achieve much better results than other baselines, and the traditional methods of word representations, such as TF-IDF, could not break a limit even with the use of the latest classification boosting models (such as XGBoost). Second, our model with 110M parameters and 8 layers was able to outperform XLNet, a much stronger and larger language model with 340M parameters and 24 layers. The results confirm that our hypothesis and method to incorporate the structure of humor is valid, as it is the only differentiating factor between the two approaches. Only using BERT has resulted in weaker result compared to XLNet, and therefore, removed from the baselines in the first evaluation.

In addition, by reviewing the wrong predictions by the trained ColBERT classifier, it becomes evident that those instances are generally close to (or even a part of) the items of the other class. For example, our model mislabeled the following two sentences as non-humor:

- "A recent study by UN has found Dexter to be the no 1 cause for ocean pollution"
- "One out of five dentists has the courage to speak their own mind"

In contrast, the following news articles were mislabeled as humor:

- "If we treated men like we do women, would they cry more at work?" (News story:[11])
- "How do we keep alias generation off facebook? permanent mittens". (News story:[12])

Even though the above examples were considered as wrong predictions, a deeper case-by-case analysis shows that the model is performing well, by being able to effectively recognize and differentiate between the two classes.

For example, the first sentence about Dexter being the cause of ocean pollution could be seen as humorous if we know the subject (character in a TV series). Otherwise, it could be seen as a factual statement and not intended to be humorous at all. Similarly, the second sentence about dentists could be interpreted as a joke about the perceived lack of courage among dentists, or it could be seen as a factual statement about the profession.

Similarly, the news articles about men crying and mittens on Facebook could be seen as humorous as they contain elements of satire or irony. However, they could also be seen as serious articles discussing gender roles and internet privacy.

This demonstrates the subjectivity of determining humor in texts, the same way that humans may find it difficult to label some of the examples. It is not uncommon for natural language data, especially in subjective areas such as humor, to contain examples that fall into a "grey area" between the classes and thus, using aggregated scores by human annotators (as done in the next evaluation) has become a common approach.

### 5.2. Evaluation on informal Spanish tweets

In the second part of the evaluation, we test the robustness of the proposed method by applying it to a new context. For this step,

[11] https://www.huffpost.com/entry/if-we-treated-men-like-women-would-they-cry-more-at_b_5904e9ace4b084f59b49f99b
[12] https://www.huffpost.com/entry/facebook-and-kids_b_1579207

our team participated in a recent shared-task with the objective to detect and rate humor of Spanish tweets (HAHA 2021[13]) using the proposed method. Here, the proposed model competes against specific solutions for the task (trained and supported by real teams of machine learning engineers) on a previously unseen dataset. The competition was organized as a part of the IberLEF 2021 forum via the CodaLab platform and attracted seventeen active teams (Chiruzzo et al., 2021).

The new setting is different from the previous evaluation for the following reasons:

1. The new task is to **detect and rate** humor in **Spanish** language, which we have no linguistic knowledge of.
2. The texts are **informal tweets**, differing in size and structure from the texts of the previous benchmark.

The organizers provided a corpus of crowd-annotated tweets using a voting scheme with six options (Chiruzzo, Castro, & Rosá, 2020): the tweet is not humorous, or the tweet is humorous and a score is given between one (not funny) to five (excellent). Data contains 36,000 tweets separated into training (24,000 tweets), development (6,000 tweets), and testing (6,000 tweets).

To predict results, we did not change the model structure, hyperparameters, or any of the pre-processing functions. However, in order to extract sentence embedding for Spanish texts, the selected BERT model is substituted from the English BERT-base-uncased to a recent Spanish equivalent (BETO-uncased (Cañete et al., 2020)). This was a required and logical step to achieve meaningful embeddings.

#### 5.2.1. Results and discussion

The contestants used a variety of models such as fastText, autoML, Spanish BERT, BiLSTM, CNN, Multilingual BERT, and LSTM. Based on the official results reported by the organizers (Chiruzzo et al., 2021), the proposed model performed strongly and achieved the 2nd place for rating humor with a 0.6246 root mean square error (RMSE). In the binary task of humor detection, our model achieved the 3rd place with a 0.8696 F1 score. The results indicate the robustness and stability of the proposed method in detecting humor in any given text.

The full description of our approach, as well as the competition criteria and detailed results and discussions for this part are presented at Annamoradnejad and Zoghi (2021).

### 5.3. Limitations and future work

This section addresses the inherent limitations of the current study while delineating promising directions for future research.

**Labeling accuracy of the novel dataset:** an automated way was used to curate the accompanying dataset and classify the instances into two categories. As seen in the evaluations, some news headlines can be considered humorous texts (or vise versa), which could lead to training problems. While the proposed method is evaluated on a second dataset for the purposes of this paper, the ColBERT dataset could be improved by human annotation to be better suited for the evaluation of future works.

**Punchline detection and continuous texts:** The current work determines the existence of humor in a given text as a whole. While this has its own applications in the classification of short text posts and user commands, in many situations the text is continuous and there are no clear cut boundaries between the sentences. In those situations, it would become essential to pinpoint the parts of speech that contain humor or separate sentences based on context.

**Multimodal humor detection:** Some recent works focused on detecting humor in voice, videos, and pictures (e.g. Choube and Soleymani (2020), Hasan et al. (2019), Patro et al. (2021), Wu et al. (2021)). While the focus of the current work is on textual content, it would be possible to apply the underlying idea of the proposed method to other types of media.

**Other humor theories:** The proposed approach was designed based on the Incongruity Theory of humor (specifically, SSTH). There are other theories that discuss the cause of laughter in humans. It would be valuable to create models based on alternative theories and compare the results.

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

## 6. Conclusion

This work introduced a novel approach based on the widely accepted Semantic Script theory of humor for the task of humor detection and rating. The technical approach utilizes BERT sentence embedding in parallel latent columns of neural networks. Furthermore, a novel dataset consisting of 200k short texts is curated and published for the task of humor detection, which mitigates the issues of existing ones regarding sample size and incompatible lexical similarity of the classes.

Based on evaluation on two different settings, our method obtained F1 scores of 0.982 and 0.869 and outperformed several state-of-the-art models. The results indicate that achieving high accuracy in the current task relies on two factors: 1) utilizing sentence embeddings and 2) incorporating the linguistic structure of humor into the design of the proposed model. While other BERT-based models also displayed strong performance compared to tree-based and other baseline models, our model that incorporated the base BERT model outperformed even larger BERT-based models (e.g. XLNet, Multilingual BERT) as well as other strong models (e.g. XGBoost, autoML, LSTM). These results show that our hypothesis and process in incorporating the structure of humor is valid.

Future work can adapt the method for developing systems for other tasks of computational humor, or assess the efficacy of the proposed parallel neural network across a broader spectrum of text classification challenges. These contributions hold the promise of pushing the boundaries of humor detection and fostering continued advancements in the field.

## CRediT authorship contribution statement

**Issa Annamoradnejad:** Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Gohar Zoghi:** Methodology, Validation, Data curation, Writing – original draft.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The novel dataset is available at our GitHub: https://github.com/Moradnejad/ColBERT-Using-BERT-Sentence-Embedding-for-Humor-Detection.

## References

Annamoradnejad, Issa, & Zoghi, Gohar (2021). ColBERT at HAHA 2021: Parallel neural networks for rating humor in spanish tweets.. In *IberLEF@ SEPLN* (pp. 860–866).
Attardo, Salvatore (2010). *Linguistic theories of humor*: *vol. 1*, Walter de Gruyter.
Bertero, Dario, & Fung, Pascale (2016). Deep learning of audio and language features for humor prediction. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 496–501).
Cañete, José, Chaperon, Gabriel, Fuentes, Rodrigo, Ho, Jou-Hui, Kang, Hojin, & Pérez, Jorge (2020). Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020* (pp. 1–20).

Castro, Santiago, Cubero, Matías, Garat, Diego, & Moncecchi, Guillermo (2016). Is this a joke? detecting humor in spanish tweets. In *Ibero-American conference on artificial intelligence* (pp. 139–150). Springer.

Chen, Tianqi, & Guestrin, Carlos (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, Lei, & Lee, Chungmin (2017). Predicting audience's laughter during presentations using convolutional neural network. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 86–90).

Chen, Peng-Yu, & Soo, Von-Wun (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).

Chiruzzo, Luis, Castro, S., Etcheverry, Mathias, Garat, Diego, Prada, Juan José, & Rosá, Aiala (2019). 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* (pp. 133–144).

Chiruzzo, Luis, Castro, Santiago, Góngora, Santiago, Rosá, Aiala, Meaney, J. A., & Mihalcea, Rada (2021). Overview of HAHA at IberLEF 2021: Detecting, rating and analyzing humor in Spanish. *Procesamiento del Lenguaje Natural*, *67*, 257–268.

Chiruzzo, Luis, Castro, Santiago, & Rosá, Aiala (2020). HAHA 2019 dataset: A corpus for humor analysis in spanish. In *Proceedings of the 12th language resources and evaluation conference* (pp. 5106–5112).

Choube, Akshat, & Soleymani, Mohammad (2020). Punchline detection using context-aware hierarchical multimodal fusion. In *Proceedings of the 2020 international conference on multimodal interaction* (pp. 675–679).

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Eysenck, Hans Jürgen (1942). The appreciation of humour: an experimental and theoretical study 1. *British Journal of Psychology. General Section*, *32*(4), 295–309.

Frenda, Simona, Cignarella, Alessandra Teresa, Basile, Valerio, Bosco, Cristina, Patti, Viviana, & Rosso, Paolo (2022). The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, *193*, Article 116398.

Giudice, Valentino (2019). Aspie96 at HAHA (iberlef 2019): Humor detection in spanish tweets with character-level convolutional RNN. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2019). CEUR workshop proceedings* (pp. 1–11).

Hasan, Md Kamrul, Rahman, Wasifur, Zadeh, Amir, Zhong, Jianyuan, Tanveer, Md Iftekhar, Morency, Louis-Philippe, et al. (2019). UR-FUNNY: A multimodal language dataset for understanding humor. arXiv preprint arXiv:1904.06618.

Ismailov, Adilzhan (2019). 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* (pp. 1–7).

Kant, Immanuel (1913). *Kritik der urteilskraft*: *vol. 39*, Meiner.

Khandelwal, Ankush, Swami, Sahil, Akhtar, Syed S., & Shrivastava, Manish (2018). Humor detection in english-hindi code-mixed social media content: Corpus and baseline system. arXiv preprint arXiv:1806.05513.

Kline, Linus W. (1907). The psychology of humor. *The American Journal of Psychology*, 421–441.

Lan, Zhenzhong, Chen, Mingda, Goodman, Sebastian, Gimpel, Kevin, Sharma, Piyush, & Soricut, Radu (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, et al. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Low, Jwen Fai, Fung, Benjamin C. M., Iqbal, Farkhund, & Huang, Shih-Chia (2022). Distinguishing between fake news and satire with transformers. *Expert Systems with Applications*, *187*, Article 115824.

Meaney, J. A. (2020). Crossing the line: Where do demographic variables fit into humor detection? In *Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop* (pp. 176–181).

Mihalcea, Rada, & Strapparava, Carlo (2005). 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). Association for Computational Linguistics.

Misra, Rishabh (2018). News category dataset.

Niculescu, Andreea, van Dijk, Betsy, Nijholt, Anton, Li, Haizhou, & See, Swee Lan (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *International Journal of Social Robotics*, *5*(2), 171–191.

Patro, Badri N., Lunayach, Mayank, Srivastava, Deepankar, Singh, Hunar, Namboodiri, Vinay P., et al. (2021). Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 576–585).

Peyrard, Maxime, Borges, Beatriz, Gligorić, Kristina, & West, Robert (2021). Laughing heads: Can transformers detect what makes a sentence funny? arXiv preprint arXiv:2105.09142.

Purandare, Amruta, & Litman, Diane (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* (pp. 208–215).

Raskin, Victor (2012). *Semantic mechanisms of humor*: *vol. 24*, Springer Science & Business Media.

Scheel, Tabea (2017). Definitions, theories, and measurement of humor. In *Humor at work in teams, leadership, negotiations, learning and health* (pp. 9–29). Springer.

Song, Yan-Yan, & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130.

Suls, Jerry M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, *1*, 81–100.

Sun, Chen, Myers, Austin, Vondrick, Carl, Murphy, Kevin, & Schmid, Cordelia (2019). Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 7464–7473).

Taylor, Julia M., & Mazlack, Lawrence J. (2004). Computationally recognizing wordplay in jokes. In *Proceedings of the annual meeting of the cognitive science society, vol. 26* (pp. 1–9).

Wang, Minghan, Yang, Hao, Qin, Ying, Sun, Shiliang, & Deng, Yao (2020). Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd annual conference of the European association for machine translation* (pp. 53–59).

Weller, Orion, & Seppi, Kevin (2019). Humor detection: A transformer gets the last laugh. arXiv preprint arXiv:1909.00252.

Wolff, Harold A., Smith, Carl E., & Murray, Henry A. (1934). The psychology of humor. *The Journal of Abnormal and Social Psychology*, *28*(4), 341.

Wu, Jiaming, Lin, Hongfei, Yang, Liang, & Xu, Bo (2021). MUMOR: A multimodal dataset for humor detection in conversations. In *CCF international conference on natural language processing and Chinese computing* (pp. 619–627). Springer.

Yang, Zhilin, Dai, Zihang, Yang, Yiming, Carbonell, Jaime, Salakhutdinov, Russ R., & Le, Quoc V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems* (pp. 5753–5763).

Yang, Zixiaofan, Hu, Bingyan, & Hirschberg, Julia (2019). Predicting humor by learning from time-aligned comments. In *Proc. interspeech 2019* (pp. 496–500).

Yang, Diyi, Lavie, Alon, Dyer, Chris, & Hovy, Eduard (2015). Humor recognition and humor anchor extraction. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2367–2376).

Zhang, Renxian, & Liu, Naishi (2014). Recognizing humor on twitter. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management* (pp. 889–898).

Zhu, Yukun, Kiros, Ryan, Zemel, Rich, Salakhutdinov, Ruslan, Urtasun, Raquel, Torralba, Antonio, et al. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision* (pp. 19–27).

Ziser, Yftah, Kravi, Elad, & Carmel, David (2020). Humor detection in product question answering systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 519–528).