Fine tuning of Language models for automation of Humor Detection

HEMANT PALIVELA[1]
TAVISHEE CHAUHAN[2]

Senior Research Scientist, Data Science, ATCI, Accenture Services, Mumbai, India
Department of Computer Engineering, Pune Institute of Computer Technology, Pune University, India
[1]hemant.datascience@gmail.com
[2]chauhantavi@gmail.com

Abstract. In this paper, we propose a method that showcases a novel approach for humor identification using ALBERT and automation of best fit loss function identification and also the Optimiser identification. We have used two configurations of ALBERT, Albert-base and Albert-large. Using different hyper-parameters, we compare their results to obtain the best results for the binary classification problem of detecting texts that are humorous and those that are not humorous. We also determine the best optimizer and loss function that can be used to achieve state-of-the-art performance. The proposed system has been evaluated using metrics that include accuracy, precision, recall, F1-score, and the amount of time required. Among multiple loss functions, Adafactor on Albert-base model have shown promising results with 99% of accuracy. Paper also talks about comparison of the proposed approach with other language models like BERT, ROBERTa to see a steep decline of 1/3rd in the time taken to train the model on 160K sentences.

Keywords: Albert, Optimiser, Natural Language Generation, Language Model, Humor Detection, Transformer

(Received November 7th 2021 / Accepted November 27th 2021)

1 Introduction

Humor is an artefact that has an intrinsic birth into each one of us. Self-obsessed or self-centred, many of insane traits, defeated, motivated are a few forms of humor. Humor might even be creative if the text and visual elements of the meme are concatenated. As in [24] a creative form of linguistic humor is used for word plays on internet meme. Strong inferential strategies are required to illuminate into the thick and multiple coatings of meaning that are implanted into the meme. The paper [23] acquires few personality traits and comparative analysis of humor style. Each one of us get affected by interpersonal skills. Subfactor studies revealed that aggressive and self-defeating humour were most closely linked to impulsivity and entitlement, whereas dominance levels influenced the use of humour to cope with stress [16]. Also, the insane component of cold-heartedness was found to be particularly empty of humour. Humor can be subjective at times. With people being more and more addicted to technology and active on various social media platforms [2, 1, 21, 20, 10], abstracts of sarcasm, humor, irony, etc. can be found in every other text sent. Considering the present covid-19 situation and its impact on human behavior. Authors [3] show the impact of laughter therapy in this stressful situation to calm down the emotions and stay healthy. Laughter therapy can be performed as an effective additional therapy to relieve the mental health burden during the COVID-19 epidemic. Humor is widely employed in language expressions as a solvent for everyday living. It is generally caused by imaginative situation or misinterpretations of unclear language. As a result, in the field

of natural language processing, identifying and recognising the humour conveyed in text is a fascinating and demanding study subject. To precisely detect humor, the authors [11] have proposed an internal-external attention neural network (IEANN). It includes working on two kinds of methods. One is the inconsistency in the humor text, and second is the obscurity in the humor text.

The remaining sections of the paper are carried out as follows: Section 2, tells about the literature review of various articles relating to humor detection. In Section 3, a detailed view of used dataset is given along with introduction to the baseline models. The next Section 4, describes about the proposed model definition. In Section 5, the experimental setup and obtained results have been shown. Also, a comparative analysis on multiple humor detection model have been done. Lastly, in Section 6, we give insights on the conclusion.

2 Literature Review

This section sheds light on the recent works that have been done relating to humor understanding, various techniques used. Humor can be comical, funny or even dark sometimes. There are different types of humor and the views of a person who finds something funny might be different from the one who doesnât find the same thing funny. It all depends on the sense of humor of a person. It can be very subjective at times. The challenge of actually being able to detect humor and the benefits of humor detection in a text has been the key factor for the development of automatic humor detection models In [8], a humor classification as well as multi modal sarcasm detection is proposed fir a Hindi-English code dataset. For evaluation of dataset, a hierarchical framework is proposed. And to understand the sequence of specific words, long short-term memory has been used (LSTM). Another author [13] suggested, humor theory in words of error detection. They intend to share that our sense of humour is acutely aware of our flaws. In this paper [25], a transfer learning framework is used for humor detection using a proposed unified multilingual model. The model is based on a multilingual BERT that has been pre-trained and can thus make predictions on corpora in Chinese, Russian, and Spanish. On the basis of a punchline of a joke, the authors have also made it possible to measure the semantic discrepancy of the setup. The challenge of actually being able to detect humor and the benefits of humor detection in a text has been the key factor for the development of automatic humor detection models [28, 30, 17]

In the past 7-8 years a lot of scientists have come up with various techniques and even implemented a lot of them in our daily used products. Google Assistant, Siri, Alexa are the key examples for the same. The challenge of classifying comedy is difficult since humour varies by culture, which means that various ethnicities see jokes differently. In chatbots and personal assistants, automatic humour detection in messages has fascinating applications. Injecting humour into computer-generated responses would be a more sophisticated consequence, making interactions more engaging and intriguing. Several individuals utilise social networking websites like Facebook, Reddit and Twitter to employ these kinds of technologies for real-time analysis of many Natural Language Emotions like comedy, sarcasm, and violence. Doing this, made the human-to-machine experience fun and enjoyable to a huge extent. But they are only correct up to some extent. Thereâs still a lot of room for improvement. In [22], the authors have proposed a convolution neural network (CNN) and bidirectional long-short term memory (biLSTM) (with and without Attention) models which takes the bilingual text as input. This can be implemented by involving more features and better computational methods in the language models. Another author [32], developed a question-answering was developed for sellers and their system. Deep learning framework was used to detect the humorous questions in the PQA system.In [18] a deep bi-directional transformer encoder was developed to score the funniest tweets. A classifier was trained which outperformed in detecting the humorous and non-humorous questions.This will help the machines to understand the context as well as the intent of the presented texts. Human Computer Interaction (HCI) is an important part of this generation. Computational humor is a subset of computational humour generation and computational humour detection [15, 19]. Many social media networks employ this to promote user retention and improve their overall service. ALBERT is an open-source implementation of TensorFlow. ALBERT has a lot of pre-trained language representation models open-sourced for better reach. Albert has multiple configurations such as base, large, extra large as well as XX large.

To fully get a sentence's comedy, it's sometimes necessary to have a lot of outside knowledge. Anecdotes, fantasy, insult, sarcasm, jokes, quotation, self-depression, and other sorts of comedy exist. Almost all of the time, there are multiple meanings hidden within a statement, each of which is perceived differently by different people, making the work of determining humour complex. As a result, we'll need an embedding that can capture the text's semantic as well as contextualised meaning.

Albert configurations	Repeating layers	Embeddings	Hidden layers	Heads	Parameters
Base-v1	12	128	768	12	11M
Large-v1	24	128	1024	16	17M
Base-v2	12	128	768	12	11M
Large-v2	24	128	1024	16	17M

Table 1: Configurations of ALBERT

3 Experiments

3.1 Dataset

We have used the dataset that is developed by [4]. The dataset consists of 200000 texts out of which 100k are humorous and the other 100k are not. Out of the many pre-trained models available, we have used Albert-base and Albert-large for our comparisons. From the pre-trained model, we extract the tokenizer. For this case, we use Albert Tokenizer. Then the token ids are assigned to the given tokens in the dataset. For example: Sentence: Ramesh is a good boy. Token: ['Rameshâ, 'isâ, âa', 'good', 'boyâ, â.â] Token id: [1802, 4458, 219, 1761, 70] We then separate the token ids and tokens, here, ('[is]', 4458). We then Encode the tokens and return pytorch tensors which return a dictionary with input ids and attention mask. The max length is set to 40.

We used a dataset that includes 200k short texts out of which 100k text are humorous and the remaining are not humorous. We conducted various experiments by varying the hyperparameters and compare the accuracy results on both the models, base cased and base uncased. With the rising demand of automated contextual understanding, many researchers have worked on humor generation techniques like Joke Analysis and Production Engine.

3.2 Introduction to baseline models and feature extraction

In this paper, we have tried to do a comparative study on various parameters of albert base and albert large respectively. The baseline models use the following standard features: In this paper, we have used Albert-base as well as Albert-large. Albert has been known to increase the training speed as that of BERT and also minimize the memory consumption, thus giving better results. Albert can be found in two versions so far; v1 and v2. Albertâs base v1 model contains 12 repeating layers, 128 embedding, 768-hidden layers, 12-heads and 11M parameters. Whereas base-v2 contains 12 repeating layers, 128 embedding, 768-hidden, 12-heads and 11M parameters. It also provides no dropouts with increased training data and longer training for better results. As for the Albert large v1 model, 24 repeating layers, 128 embedding, 1024-hidden, 16-heads and

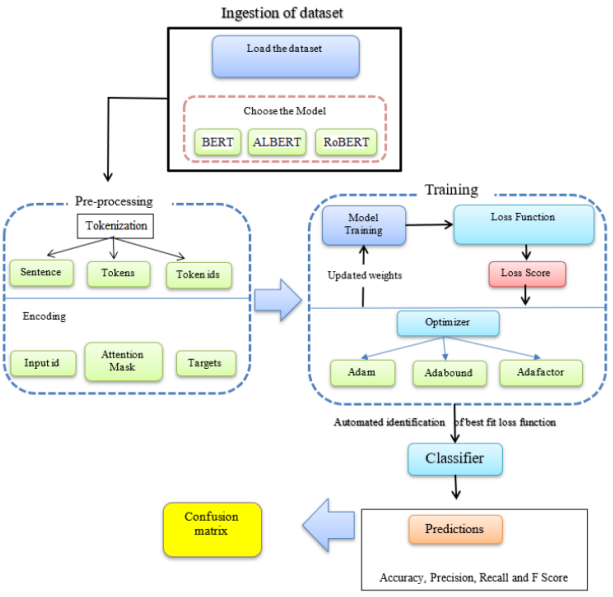
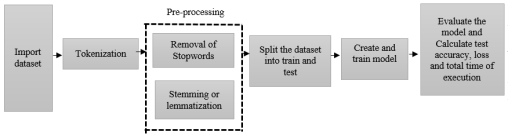


Figure 1: Architecture/Flow for identifying if a text is humorus

17M parameters are obtained and large-v2 has the same parameters with less or no dropouts. ALBERT is an unsupervised language representation learning algorithm [14]. From the comparison, it is visible that version v2 is the better choice. Adam makes use of alpha parameter which stores the learning rate, beta stands for the weight decay and epsilon is used to prevent the weight from being zero as that would interrupt the training process immediately.

4 Model definition

In this paper, we showcase the existing ALBERT model by fine tuning it so as to make the model master in classifying text that are humorous and those that are not. A lot of studies on the same topic has been done with datasets like tweets on twitter, reddit posts, various language posts, even on some text messages. As humor is a lot of times very subjective to the person as well as the situation, it is not always easily understandable. Even by people. In this paper, we have tried to make it simpler by using the Albert Language model. When given a text or tweet, the system is supposed to acknowledge if the text was a joke (humorous) or not. We also observe the difference between accuracy rate, precision, F1 score, etc. when given two different parameters on Albert base as well as Albert large. In our proposed model, we demonstrate how can humor be detected using the existing ALBERT base and ALBERT large models and fine tune them. Our Proposed model shown in 1 helps to automate right language model,



(2).PNG

Figure 2: The processing that takes place in proposed model

maps the right loss functions and includes right optimisation for the problem to be solved.

4.1 Fine tuning

Fine tuning is the most essential step throughout the process of building the model to detect humor in text. As we have used pre-trained models that is Albert base and Albert large, we need to fine tune these models in accordance to the new specialized features of the dataset. We need to make the pre-trained model learn about the specialized features of the current dataset, and fine tune them with the already existing features.

Figure 1 and 2 explains the working of our proposed model.

5 Model evaluation

5.1 Experimental setup

In this model, we made use of various machine learning libraries like pandas, sklearn, matplotlib, numpy as well as transformers.[7, 29]. The data is split into training and testing sets. We then extract embeddings for the text from the pre-trained model. The data is then sent to the classifier. The humor classifier has two classes âTrueâ and âFalseâ. The batch size of input ids and attention mask is multiplied by sequence length. The amount of weights to be updated while training the model is known as the learning rate. It can be the most important hyperparameter at times. We have used 1e-5 to fine tune our model. The default learning rate [24, 25, 28] for SGD is 0.001. SGD also provides different adaptive learning rate optimizers such as Adam, RMS, Adagrad, etc. The optimizers used for this model are Adam, AdaBound and Ada Factor with the learning rate set to 1e-5 as these parameters are bound to give the best accuracy and precision rates. Adaptive learning rates are used to increase the training rate all the while minimizing the overall time consumption.

From the large variety of loss function options available, such as hinge loss, square hinged loss, Cross entropy, etc. The loss function we use for this model is Cross entropy as this is a binary classification problem. Cross entropy is built on the idea that a specific number of pieces are required to make a comparison like structure between one distribution to another, in this case between the humorous texts and non-humorous oneâs. It helps answer the question âIs this text funny?â with a Yes or No. The training loss, testing loss, time taken to complete the training are taken into account for the comparison. As the projected likelihood differs from the actual label, cross-entropy loss grows.

For number of classes C=2, cross entropy can be calculated as:

$$H = -(ylog(p) + (1-y)log(1-p)) \quad (1)$$

Similarly, for C>2, cross entropy will be as follows:

$$H = -\sum_{c=0}^{M} y_{o,c} log(p_{o,c}) \quad (2)$$

where H is the cross entropy function, y is the binary indicator, and p is predicted probability observation of that particular class. And o stands for observations and c indicated the respective class

Albert-base usually takes about 43 minutes on average to run the model on different parameters, whereas Albert-large takes about 57-58 minutes for the same as it trains on more parameters compared to the base configuration. Keeping in mind the train/val accuracy, we have made sure to anneal the learning rate over time. We have made use of a linear scheduler for this purpose with relevant weight decays.[9, 5, 31]

The model is trained on 3 Epochs for higher accuracy rate. The prediction part includes classifying a text to be humorous or not. This includes the precision, F1 scores and recall values. The predictions can be classified into two classes: True predictions and False predictions. Then a confusion matrix of Predicted Humor VS True Humor is showcased predicting the True positive, False positive, True negative and False negative is presented for a clear understanding.

The training of the proposed model has been done on Windows machine using an AMD Threadripper processor of 64 cores, 128 threads and 256 MB cache. The graphic card used was RTX3080 and RAM of about 128GB.

5.2 Experimental results

The calculations are done on the batch size 64 and 3 epochs. By fluctuating the loss functions of Adam, Ada Factor and AdaBound between albert-base and albert-large. Stochastic Gradient Descent has been tweaked to create the loss function Adam. The update rule of Adam is simply a modified version of the update rule

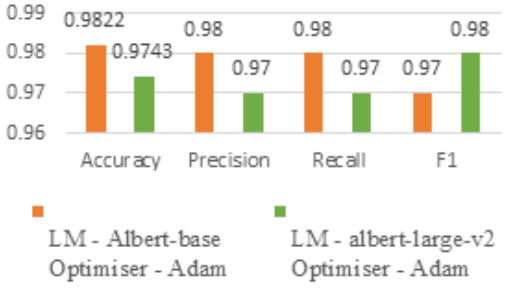


Figure 3: Using Adam Optimizer

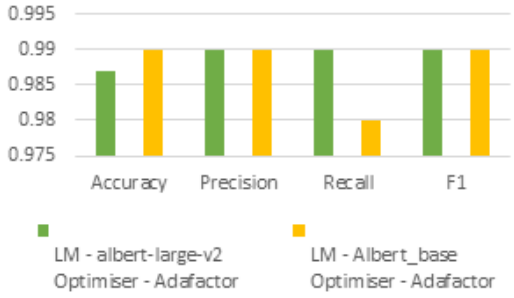


Figure 4: Using AdaBound Optimizer

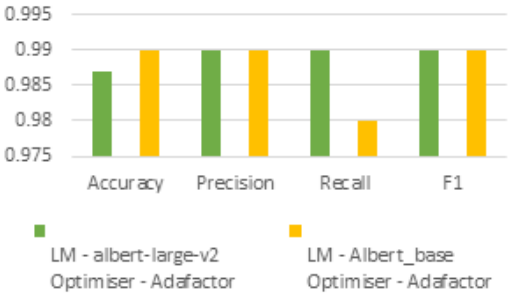
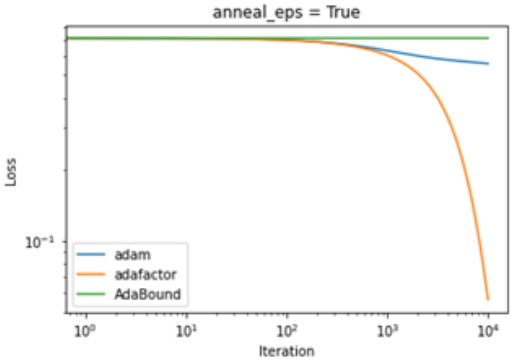


Figure 5: Using AdaFactor Optimizer



(1).png

Figure 6: Comparison of loss details for all the three optimisers

Table 2: Detailed comparison of loss functions with language models

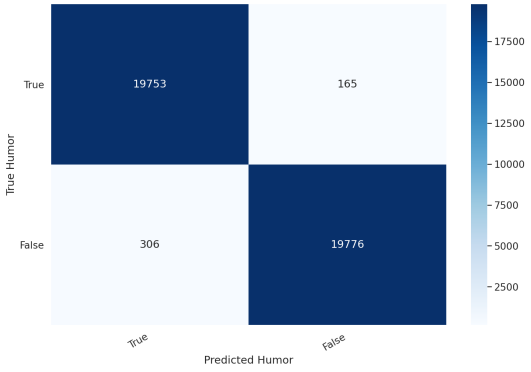
	Accuracy	Precision	Recall	F1-score	Time require
Loss Function-AdaFactor Model-Albert base	2***99**	2***99**	2***98**	2***99**	2***43minutes**
Loss Function-AdaFactor Model-Albert large	2*98.7	2*99	2*99	2*99	2*57minutes
Loss Function-Adabound Model-Albert base	2*98	2*97	2*99	2*98	2*29min33sec
Loss Function-Adabound Model-Albert large	2*98.7	2*99	2*99	2*99	2*57minutes
Loss Function-Adam Model-Albert base	2*98	2*99	2*98	2*97	2*43min22sec
Loss Function:Adam Model-Albert large	2*97.4	2*97	2*97	2*98	2*58minutes

of Gradient descent. In Adam, the parameter vector (θ) is subtracted from the EMA of the first moment of the gradient scaled by the square root of the second moment of the moment.

$$\theta_{t+1} = \left(\theta_t - \frac{n}{\sqrt{v_t}} + \epsilon\right) m_t \tag{3}$$

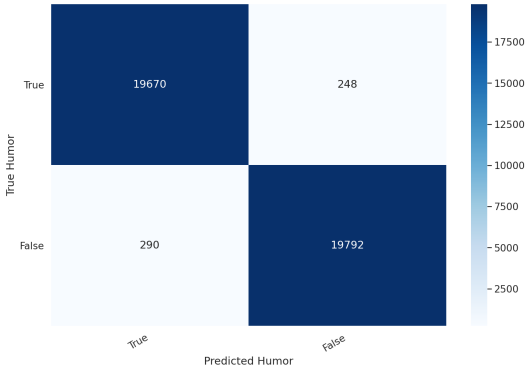
We find that The base model gives 98% accuracy, 99% precision, 98% recall and 97% F1 score in 43 minutes and 22 seconds whereas the albert-large-v2 provides the 97.4% accuracy and 97% precision, 97% Recall and 98% F1 score in 58 minutes with Adam loss function (as shown in Fig 2). With Ada Factor, the base gives 99% accuracy, 99% precision,98% recall and 99% F1 score in 43 minutes whereas large gives 98.7% accuracy, 99%precision and 99% recall; 99% F1 score in 57 minutes (as shown in Fig 3). With Ada Bound, the base gives 98% accuracy, 97% precision,99% recall and 98% F1 score in 29 minutes 33seconds whereas large gives 98.7% accuracy, 99%precision and 99% recall; 99% F1 score in 57 minutes (as shown in Fig 4).

The above table shows the details about precise selection of loss function with appropriate results.



(2).png

Figure 7: Confusion matrix for BERT



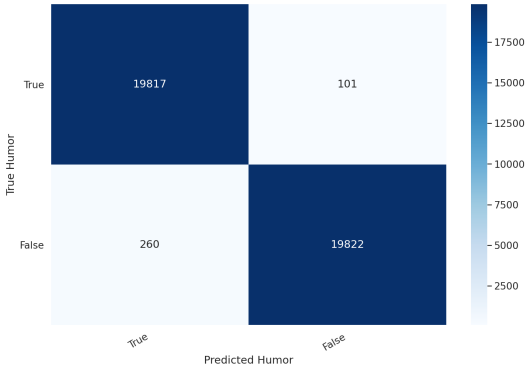
(2).png

Figure 8: Confusion matrix for ALBERT

The figure, graph chart can be written as per given above schedule. Instead of the classical approach [6] to use Stocastic Gradient Descent, we have made use of Adam optimizer to iteratively update the network weights. Instead of keepiing the same learning rate throughout the model, Adam provides adaptive rate depending on each fold. Adam takes into account uncentered varriance to provide the best weights possible. AdaBound is a mixture of Adam and AMSGrad. In AdaBound, the lower and upper bound are initially taken as zero and infinity to smoothly reach a constant final step size as opposed to AMSGard, this eventually fine tunes to the SGD with the increasing time steps. This is actually able to perform well on large datasets as well. Thus, provides similar accuracy to Adam. The confusion matrix for the models bert, Albert and roberta is shown is the figures Figure 7, Figure 8 and Figure 9

Table 3: Comparison with other research work

Sr.No	Title	Accuracy	Year
1	Multi-modal sarcasm Detection and humor classification in code-mixed conversations	83%	2021
2	Humor detection via an internal and external neural network	93.88%	2020
3	Humor Knowledge Enriched Transformer for Understanding Multimodal Humor	79.41%	2021
4	Deep Learning Techniques for Humor Detection in Hindi-English Code-Mixed Tweets	73.80%	2019
5	Humor Detection in Product Question Answering Systems	90.80%	2020
6	A BERT-based Approach for Automatic Humor Detection and Scoring	91.00%	2019
7	Dutch Humor Detection by Generating Negative Examples	98.80%	2020
8	Applying a Pretrained Language Model to Spanish Twitter Humor Prediction	84.58%	2019
9	ColBERT: Using BERT Sentence Embedding for Humor Detection	98.20%	2020
10	Humor Detection: A Transformer Gets the Last Laugh	93.00%	2019
11	**[Our Work]Automation and Fine-Tuning Hyper-Parameters for Humor Detection using Language Models**	**99%**	



(2).png

Figure 9: Confusion matrix for ROBERTA

5.3 Comparative analysis

The following table 2 shows a comparative analysis of humor detection based related findings proposed by multiple authors.

In Sr.no.1, author developed a multi-modal sarcasm detection and humor classification in conversational dialog. In Sr.no.2, the authors have proposed an internal-external attention neural network (IEANN). It includes working on two kinds of methods. One is the inconsistency in the humor text, and second is the obscurity in the humor text. As a result, the proposed model has comparatively achieved the state-of-art performance. The authors in Sr.no.3 tells that using external knowledge a multimodal called as Humor Knowledge enriched Transformer (HKT) has been proposed to capture the humor expressions. For punchline URFUNNY, the proposed system achieves 77.36%. Where as in Sr.no.4 authors have proposed a convolution neural network (CNN) and bidirectional long-short term memory (biLSTM) (with and without Attention) models which takes the bilingual text as input. The results stated that the proposed model outperformed by 73.6%. It showed an improvement of 4% as compared to other state-of-art models. In Sr.no.5, a question-answering was developed for sellers and their system. Deep learning framework was used to detect the humorous questions in the PQA system. A classifier was trained which outperformed in detecting the humorous and non-humorous questions. In Sr.no.6, A deep bi-directional transformer encoder was developed to score the funniest tweets. The first test achieved 0.784 F1-score, where as task 2 showed RMSE of 0.910. In Sr.no.7, a RobBERT model is introduced and proposed for text creation algorithms to resemble the original joke collection, rather than using wholly distinct non-humorous texts, to raise the challenge of the learning method. Even though other models performed okay when non-jokes came from completely other domains, RobBERT is the one that could make a distinction between jokes and denial or obstructive examples [27]. In Sr.no.8, the language model was trained from start centred on Spanish based on the twitter corpus which then was able to give insights to proposed model [12]. In Sr.no.9, BERT was used in the proposed method to build the embeddings for phrases in the text document, which were then feeded to the neural network [4]. In Sr.no.10, A novel approach based on the transformer is introduced that is to determine if a joke is humorous or not [26]. Lastly, Sr.no.11 sheds light on our proposed system that outperforms among multiple systems.

6 Conclusion

We have done an automated selection for proposed model using multiple loss functions to achieve the highest accuracy that best suits our model. Multiple loss function we choose for automated results include Adam, Adafactor and Adabound. Among these loss functions, from table 2 it is clear that the AdaFactor optimizer on the Albert-large model provides the best accuracy, precision, recall rate as well as F1 score i.e. 99% in just 57 minutes. The proposed system achieved 99% of accuracy which has outperformed as compared to other humor detection models. We have also done a comparative study of recent findings in the topic of humor detection, which gives a glance at respective studies of multiple authors and our proposed model have outperformed within the comparative analysis.

References

[1] Affonso, E. T., Nunes, R. D., Rosa, R. L., Pivaro, G. F., and Rodriguez, D. Z. Speech quality assessment in wireless voip communication using deep belief network. *IEEE Access*, 6:77022–77032, 2018.

[2] Affonso, E. T., Rosa, R. L., and Rodriguez, D. Z. Speech quality assessment over lossy transmission channels using deep belief networks. *IEEE Signal Processing Letters*, 25(1):70–74, 2017.

[3] Akimbekov, N. S. and Razzaque, M. S. Laughter therapy: A humor-induced hormonal intervention to reduce stress and anxiety. *Current Research in Physiology*, 2021.

[4] Annamoradnejad, I. and Zoghi, G. Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*, 2020.

[5] Apicella, A., Isgrò, F., Pollastro, A., and Prevete, R. Dynamic filters in graph convolutional neural networks. *arXiv preprint arXiv:2105.10377*, 2021.

[6] Bahuleyan, H. Natural language generation with neural variational models. *arXiv preprint arXiv:1808.09012*, 2018.

[7] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., and Bengio, Y. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.

[8] Bedi, M., Kumar, S., Akhtar, M. S., and Chakraborty, T. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *arXiv preprint arXiv:2105.09984*, 2021.

[9] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[10] de Almeida, F. L., Rosa, R. L., and Rodriguez, D. Z. Voice quality assessment in communication services using deep learning. In *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pages 1–6. IEEE, 2018.

[11] Fan, X., Lin, H., Yang, L., Diao, Y., Shen, C., Chu, Y., and Zou, Y. Humor detection via an internal and external neural network. *Neurocomputing*, 394:105–111, 2020.

[12] Farzin, B., Czapla, P., and Howard, J. Applying a pre-trained language model to spanish twitter humor prediction. *arXiv preprint arXiv:1907.03187*, 2019.

[13] Kramer, C. A. I laugh because it's absurd: Humor as error detection. 2021.

[14] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[15] Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

[16] Lobbestael, J. and Freund, V. L. Humor in dark personalities: An empirical study on the link between four humor styles and the distinct subfactors of psychopathy and narcissism. *Frontiers in psychology*, 12, 2021.

[17] Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[18] Mao, J. and Liu, W. A bert-based approach for automatic humor detection and scoring. In *IberLEF@ SEPLN*, pages 197–202, 2019.

[19] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[20] Rodriguez, D. Z. and Bressan, G. Video quality assessments on digital tv and video streaming services using objective metrics. *IEEE Latin America Transactions*, 10(1):1184–1189, 2012.

[21] Rodríguez, D. Z., Rosa, R. L., Costa, E. A., Abrahão, J., and Bressan, G. Video quality assessment in video streaming services considering user preference for video content. *IEEE Transactions on Consumer Electronics*, 60(3):436–444, 2014.

[22] Sane, S. R., Tripathi, S., Sane, K. R., and Mamidi, R. Deep learning techniques for humor detection in hindi-english code-mixed tweets. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–61, 2019.

[23] Tsukawaki, R. and Imura, T. Incremental validity of the dual self-directed humor scale in predicting psychological well-being: Beyond the big five personality traits and four humor styles. *Current Psychology*, pages 1–10, 2021.

[24] Vasquez, C. and Aslan, E. âcats be outside, how about meowâ: multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171:101–117, 2021.

[25] Wang, M., Yang, H., Qin, Y., Sun, S., and Deng, Y. 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*, pages 53–59, 2020.

[26] Weller, O. and Seppi, K. Humor detection: A transformer gets the last laugh. *arXiv preprint arXiv:1909.00252*, 2019.

[27] Winters, T. and Delobelle, P. Dutch humor detection by generating negative examples. *arXiv preprint arXiv:2010.13652*, 2020.

[28] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[29] Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[30] Zhang, R. and Liu, N. Recognizing humor on twitter. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 889–898, 2014.

[31] Zhu, T., Li, J., Liu, X., Jiang, Y., and Xia, S.-T. Attention on attention sparse dense convolutional network for financial signal processing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3300–3304. IEEE, 2021.

[32] Ziser, Y., Kravi, E., and Carmel, D. Humor detection in product question answering systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 519–528, 2020.