DilBERT^2: Humor Detection and Sentiment Analysis of Comic Texts Using Fine-Tuned BERT Models

Krzysztof Garbowicz , Lydia Y. Chen , Zilong Zhao
TU Delft

Abstract

The field of Natural Language Processing (NLP) techniques has progressed rapidly over the recent years. With new advancements in transfer learning and the creation of open-source projects like BERT, solutions and research projects emerged implementing new ideas in a variety of domains, for tasks including text classification or question answering. This research focuses on the task of humor detection and sentiment analysis of comic texts with the use of a fine-tuned BERT model. Never before has a fine-tuned BERT model been used for the task of humor detection in a text coming from an artwork. Comic text features domain-specific language, affecting the meaning and structure of commonly used grammar and vocabulary. This may differ from the language people use every day, and from the language the existing humor classifiers used for training. This research contributes to the NLP field with new models and datasets for humor detection and sentiment analysis, and reports on techniques to improve the training times and the accuracy of a pre-trained BERT model on small datasets. The proposed solution trained on comic datasets outperforms the chosen baselines and could be used as a reliable classifier in the given domain. Moreover, the results indicate that techniques reducing the training time of the model can positively contribute to its performance.

I. Introduction

The recent developments in natural language processing (NLP) techniques using pre-trained language models have outperformed previously used methods for tasks related to language understanding [1], [2]. One of the most popular and successful models created in recent years called BERT (Bidirectional Encoder Representations from Transformers) [1], introduced a technique of unsupervised pre-training of deep, contextual, and bidirectional language representations. The pre-training was done on a dataset of over 3 billion words [1], which allows for learning complex linguistic features in the training phase and adjusting the output layers afterward to achieve a network able to perform different language understanding tasks. Calibration of the output layers, also called fine-tuning, is performed on an already pre-trained language model and can be used to get predictions for text classification or question answering tasks. Fine-tuning can be performed with domain-specific, labeled data in much smaller quantities than data in the pre-training phase, and it can be done in much less time.

This research tries to answer if it is possible to fine-tune a pre-trained BERT model on a domain of comic text, on task of humor detection and sentiment analysis. The performance of the proposed models is compared to different configurations and selected baselines. Performance is evaluated on datasets from selected comics and other humor-detection studies. The further goal is to answer if the classification results can be used as a condition for comic image generation generative adversarial networks (GANs), developed concurrently with the presented study. The goal of the surrounding research is to extract images and text from *phdcomics.com* and *dilbert.com*, perform pre-processing or classification of the obtained data, and feed selected information into the GANs which would create new comic illustrations. The classification of dialogues and descriptions in the comics can deliver additional information about the topic or the sentiment of a statement. An example use of that information could be varying the background color of the generated comic based on the sentiment or the humor of the provided text.

Previous contributions include ColBERT [3], a classifier using BERT tokens trained for humor detection task, and ref. [4] which uses a fine-tuning approach. Additional research in the field of solving NLP tasks with a pre-trained BERT model at their core includes data pre-processing techniques that could significantly decrease the time of training a classifier [5]. Other focus on improving the accuracy when trained on smaller datasets [6][7].

This research combines the techniques of fine-tuning a pre-trained BERT model on a small dataset [6][7], and training the model using batches of data of dynamic size [5]. The proposed models are pre-trained BERT models with adjusted output layers. The input text is tokenized to

a format required by BERT, and fed into the model which produces a classification label. By training the models with different training parameters, on datasets from comics and previous humor-detection studies, the accuracy and F1-score are measured. The goal is to achieve a reliable humor and sentiment classifier, in the new domain of comic texts, that results could be used as one of the input conditions of the comic generating GAN. .

The report is laid out in the following structure. Section II describes the previous work in the domain of fine-tuning a BERT model and humor detection. Section III describes the creation of the datasets and the proposed models. Section IV explains how the experiments were performed, and describes the experimental setup. Section V presents the achieved results. Next, in section VI the aspects of ethics and reproducibility are reflected upon. Section VII contains the discussion and anlysis of obtained results. In section VIII conclusions and suggestions for future work are made.

II. Related Work

Over the recent years, research surrounding applying pre-trained BERT models brought novel approaches allowing for trimming the initial architecture, and faster training without losing accuracy in a given task [6][7]. Concurrently, studies in different text domains were performed to explore the possibilities of applying BERT embeddings to solve a variety of tasks like text classification [2], question answering [8], or named entity recognition (NER) [9]. The related work section describes the Bert model, humor detection methodology, existing humor detection models, and the technique of fine-tuning a pre-trained BERT model.

BERT Model The Bidirectional Encoder Representation from Transformers (BERT) [1] is a model trained in an unsupervised fashion on tasks of masked language modeling and next sentence prediction on a dataset consisting of BookCorpus [10] and English Wikipedia. Once the model was pre-trained it was distributed as an open-source project and has been one of the state-of-the-art models in NLP since. Two versions of the model exist BERT$_{LARGE}$ (24 encoder layers, 340M parameters) and BERT$_{BASE}$ (12 encoder layers, 110M parameters). The model learns the context from tokenized representations of the input text, and stores different information at different encoder layers [11]. The format of the input tokens has a specified structure. The encodings begin with a **[CLS]** token, followed by tokens representing sub-words of the input text, and might end with **[PAD]** tokens in case the encoded text is shorter than the specified batch length (max=512). By using transfer learning and attention [12], after pre-training the architecture distinguishes between complex linguistic features. That knowledge comes at a high cost of financial and computer resources. However, since pre-training needs to be done only once, models implementing different architectures or pre-training techniques, such as RoBERTa [13], TaBERT [14], CamemBERT [15], or ALBERT [16] were developed, and broadened the understanding of the architecture and capabilities of the BERT model.

	# Characters	# Words	# Unique Words	# Punctuation	# Duplicate Words	# Sentences
mean	65	12	12	2	0	1
std	22.73	4.482	4.460	1.515	0.854	0.685
min	4	1	1	0	0	1
median	65	12	12	2	0	1
max	145	30	30	19	6	5

TABLE I: General statistics of the grammatically correct Dilbert comics dataset used for the study.

Humor Detection The structure of humor chosen for this study follows The Script-based Semantic Theory of Humor (SSTH) [17]. By its main hypothesis, it can be assumed that a text is humorous if the context of its parts is compatible with each other, and when it contains a punchline, characterized by some parts being opposite to each other. The punchline could be distinguished by a change in the use of grammar, vocabulary, or punctuation. There are different theories of humor that rely on more complex data, such as monitoring human reactions or laughter. On the contrary, SSTH focuses exclusively on linguistics.

Humor Classifiers Previous studies based on a pre-trained BERT model, solving the task of humor detection, ColBERT [3] and "Humor Detection: A Transformer Gets the Last Laugh" (TGtLL) [4], achieved state-of-the-art results when trained on large, prepared datasets. ColBERT uses BERT sentence embeddings and an eight-layer classification network. It achieves state-of-the-art results, being trained on a dataset of 200K examples of humorous and nonhumorous text. TGtLL uses a fine-tuned BERT model on big collections of jokes and nonhumorous text. In contrast with previous contributions, new datasets for this study are much smaller, and consist of text scraped from popular comics.

Fine-Tuning By adjusting the output layers of pre-trained BERT models architectures were developed to solve different NLP tasks. Datasets of around 10,000 - 20,000 training examples were often used for fine-tuning. The effects of using smaller datasets, different training functions, and training parameters have been recently investigated. Studies show that a dataset of 1000 examples is often sufficient in different text classification tasks [6] [7]. Other studies focused on discovering techniques enabling a faster training process [5]. The new comic datasets for the study comprise of around 1000 training examples, and experiments with the newly proposed techniques are performed.

III. Methodology

This section explains the chosen approaches, describes the process of creating the datasets, and building the models for classification tasks.

A. Datasets

Three different comics were proposed as sources of new data for the study, *PhdComics.com dilbert.com*, and

text	humor
Are you kidding?	False
My date complained about her life all night long!	False
That is cruel and senseless. I am thoroughly ashamed of you.	False
Maybe you could support it now and then stab me in the back later.	True
How many ten dollar mouse pads can we get for 10,000? I hope this is a panic attack.	True
...All the people with excessive nose hair and anyone who insists on being called doctor.	True

Fig. 1: Samples from the dataset of Dilbert comics for humor detection with text with correct grammatical structure

text	humor
we work for the same company. my cubicle is down the hall.	False
we bought a start up just so we could get the engineers, including you.	False
want some advice? why?	False
hows this different from a layoff? with layoffs you get to keep your dignity.	True
i started an online marketplace for dumb criminals.	True
id better not keep her waiting at the door. do not anger jabba the date.	True

Fig. 2: Samples from the dataset of Dilbert comics for humor detection with text with incorrect grammatical structure

Garfield. Due to the varying form of the language used, common important connections between the text and the illustration, and difficult to scrape text data, illustrations from *PhdComics.com* were not selected to be a part of the dataset. Instead, *dilbert.com* and *Garfield* became the sources of the comic texts. Both comics exist only in a few different panel structures, and use simpler forms of texts than *PhdComics.com*. In comparison to often long and complex text from *PhdComics.com*, *dilbert.com* and *Garfield* consist mostly of short monologues, dialogues, or descriptions. Whatsmore, transcriptions from *Garfield* comics can be found online, which allowed for creating a dataset following the exact formatting as in the original comic.

Both datasets for the study consist of over 1300 texts extracted from single panes from the comics. Each text sample in the Dilbert dataset is labeled for binary classification of humor detection, and multi-label classification of sentiment analysis. Text samples from the Garfield dataset are labeled for the task of humor detection. The humor labels for the data follow the description of the SSTH [17], which was also chosen by previous studies [18]. The sentiment analysis uses three emotions to categorize the text. Negative, represented with a 0, neutral, represented with a 1, and positive, represented with a 2. Text extracted from comics contains many abbreviations and onomatopoeias, like ”pow” or ”ghaa”. The vocabulary of a pre-trained BERT model only includes real English words and names. That is why two versions of the Dilbert dataset are created to experiment with the input format. One version uses the structure of text resulting from the data scraping. The text is all lowercase, and all symbols other than punctuation are omitted. Sentences from the other

text	humor
Cats are smarter than dogs.	0
Garfield...	0
We all have times when we just can't get started.	0
A $75 fine for our ugly yard Santa. Gets you right here.	1
I am the ghost of a hamburger that you ate!	1
Good morning. The older we get, the louder we wake up.	1

Fig. 3: Samples from the dataset of Garfield comics for humor detection with text with correct grammatical structure

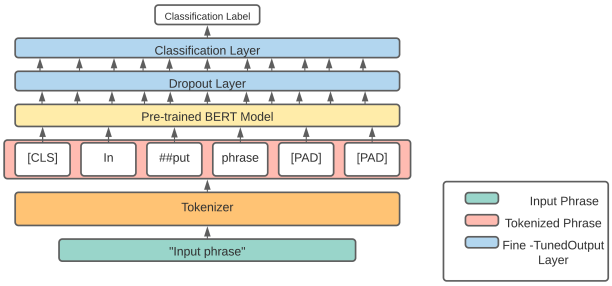

Fig. 4: The architecture of the fine-tuned models

dataset are rewritten with expanded contractions and use uppercase formatting. Table I presents general information about the grammatically correct dataset. Both datasets are used when training the model to perform experiments and observe the influence of the use of correct grammar and vocabulary.

For comparison with previous humor detection classifiers, the proposed model is evaluated on the datasets from TGtLL and ColBERT studies. Pun of The Day and Short Jokes datasets were published alongside the TGtLL study. The Pun of The Day dataset is composed of one-sentence humorous and non-humorous texts. The Short Jokes dataset consists of text reaching the maximum of couple sentences and is comprised of jokes from the Short Jokes Kaggle dataset [19], and sentences from news websites. To compare the proposed model with ColBERT, and experiment with using small datasets for fine-tuning a pre-trained BERT model, training is done on 1000 examples from the ColBERT dataset.

B. Models

Models for the study are fine-tuned pre-trained BERT models, with output layers with the number of outputs dependent on a task. The classifiers consist of a dropout and a fully connected linear layer. The proposed classifier is coined with a name *DilBERT*^{2}. During training, the data is tokenized, and batches for the training phase are organized in a fixed and a dynamic way [5] for comparison. The dynamic way of organizing data is characterized by sorting the input by its length and creating batches of input of similar size. The text within a batch is then padded to the length of the longest sequence in a batch or the chosen maximum length. The BERT model requires the data within a batch to be of the same length, however, this length can vary between batches. The dynamic approach leads to better run-time due to the lower number of padding tokens. The padding tokens do not carry the meaning of the analyzed text, but they can be seen as unnecessary use of memory when padding all the samples to the same length. The choices of the optimizer, the scheduler, and parameters were influenced by previous research [6] [7].

IV. Experimental Setup

This section describes the experimental setup and chosen baselines for the comparison of the DilBERT^2 model.

All of the experiments performed for this study have been done on the Google Colab platform, hosting a Jupyter Notebook instance with free access to several different GPUs. The code for the models was written in the PyTorch framework with the HuggingFace Transformers library providing the tokenizers and pre-trained models.
Experiments were executed on the case and uncased BERT$_{LARGE}$ and BERT$_{BASE}$ models with the size of the training batch of 16. The output layers of the developed fine-tuned models consist of a dropout and a dense output layer, with the number of output labels dependent on the task. The probability of dropout is set to $p = 0.1$. The optimizer for the model is the AdamW optimizer with a learning rate of 2e-5, weight decay λ= 0.01, and bias correction. Warmup of the learning rate lasts for the first 10% of the training steps. The datasets are split into a 75%-25% train-test split. The model is trained for a maximum of 20 epochs, and varying random seeds.

The model tasked with humor detection is compared against the following classifiers.

1) **Decision Tree**
   A classification technique based on distinguishing between different numerical features of the input, and predicting the target label by using the knowledge from the training phase when the tree was built. *CountVectorizer* is used to encode text into numerical vectors.

2) **SVM**
   The Support-Vector-Machine classifier tries to find a separation line between two sets of points representing mapped data. Label predictions of new examples are obtained by encoding the input into the corresponding space and looking at which side of the space the point falls into. *CountVectorizer* is used to map the input text into numerical representations.

3) **Multinomial naive Bayes**
   The Multinomial Naive Bayes algorithm uses the Bayes theorem on encoded text to predict the label. Text is encoded into vectors with the use of *CountVectorizer*
   .

4) **XGBoost**
   The XGBoost library provides efficient implementations of the gradient boosting technique to make predictions based on the decision tree algorithm. *CountVectorizer* is used to map the input text into numerical vectors.

5) **A Transformer Gets the Last Laugh (TGtLL)**
   One of the previous studies on the task of humor detection that published the datasets, as well as the results achieved by the proposed model. The datasets include the Pun of the Day dataset [20], composed of short phrases, and the Short Jokes dataset published alongside the study. The performance of DilBERT^2, trained in the described approaches on each of the datasets, is compared against the results of the TGtLL model.

6) **ColBERT**
   ColBERT is a state-of-the-art model in the humor detection task, scoring 98% accuracy and F1-score on its test set. Achieving such a good result could be explained by big quantities of training data (160K examples). The data, however, consists of jokes and nonhumorous texts like news headlines, which is different from the monologues and dialogues used to train the DilBERT^2 model. A pre-trained ColBERT model is compared against DilBERT^2 to assess whether it is beneficial to use the state-of-the-art model instead.

Method	Configuration	Accuracy	Precision	Recall	F1
Decision Tree		0.598	0.668	0.708	0.684
SVM	sigmoid, gamma=1.0	0.491	0.563	0.606	0.581
Multilingual NB		0.506	0.576	0.619	0.595
XGBoost		0.550	0.621	0.701	0.656
ColBERT		0.591	0.740	0.994	0.740
DilBERT^2	Base-Cased: Fixed Batching	0.790	0.853	0.787	0.819
DilBERT^2	Base-Cased: Dynamic Batching	0.799	0.829	0.810	0.819
DilBERT^2	Large-Cased	0.769	0.816	0.753	0.784

TABLE II: Comparisons of different methods on the task of humor detection on the grammatically correct version of the Dilbert dataset.

Method	Configuration	Accuracy	Precision	Recall	F1
Decision Tree		0.577	0.640	0.617	0.627
SVM	sigmoid, gamma=1.0	0.546	0.617	0.608	0.612
Multilingual NB		0.566	0.621	0.681	0.649
XGBoost		0.621	0.659	0.7077	0.682
ColBERT		0.588	0.586	0.997	0.738
DilBERT^2	Base-Uncased: Fixed Batching	0.756	0.801	0.732	0.765
DilBERT^2	Base-Uncased: Dynamic Batching	0.778	0.826	0.748	0.785
DilBERT^2	Large-Uncased	0.739	0.755	0.761	0.758

TABLE III: Comparisons of different methods on the task of humor detection on the dataset composed of parsed text from Dilbert comics.

V. Results

This section presents the results obtained from the performed experiments. All reported results of the base-model were recorded after training the model for 6 epochs, the large-models were trained for the duration of 20 epochs. The rest of the settings follow the description in the Experimental Setup section. This section covers the results of the task of humor detection, sentiment analysis, and the use of dynamic batching.

A. Humor Detection

Tables II-IV compare the performance of the chosen methods in the humor detection task on the comic text domain. DilBERT^2 significantly outperforms the chosen baselines,

Method	Configuration	Accuracy	Precision	Recall	F1
Decision Tree		0.603	0.416	0.233	0.284
SVM	sigmoid, gamma=1.0	0.594	0.427	0.332	0.363
Multilingual NB		0.632	0.481	0.236	0.303
XGBoost		0.618	0.440	0.233	0.297
DilBERTˆ2	Base-Cased: Fixed Batching	0.813	0.779	0.646	0.706
DilBERTˆ2	Base-Cased: Dynamic Batching	0.814	0.757	0.641	0.694
DilBERTˆ2	Large-Cased	0.809	0.821	0.652	0.727

TABLE IV: Comparisons of different methods on the task of humor detection on the dataset composed of text from Garfield comics.

Model	Configuration	Time	Accuracy	Precision	Recall	F1
DilBERT^2	Fixed Batching	5min 48s	0.991	0.990	0.992	0.991
DilBERT^2	Dynamic Batching	2min 15s	0.991	0.987	0.992	0.990
TGtLL			0.930	0.930	0.931	0.931

TABLE V: Comparison of DilBERT^2 against TGtLL reported performance on the Pun Of The Day dataset.

achieving scores of almost 80% when trained on the base-cased version, and a score of 76% when using the large-cased model. When the uncased version of the model is used the accuracy of DilBERT^2 decreases by 2-3%. Table IV presents the overview of achieved scores on the Garfield dataset. The chosen baselines are also inferior to DilBERT^2 in the Garfield domain. The found accuracy of DilBERT^2 in different configurations is around 80% with an F1 score of around 70%.

The results achieved by DilBERT^2 trained and evaluated on the datasets from other humor-detection studies outperform, or fall within a close margin of the previously reported scores. Table V presents the results of training DilBERT^2 on the Pun Of The Day dataset. The found accuracy and F1 score of DilBERT^2 are 99%. That is an increase of 7% over the reported results by the previously fine-tuned model. Table VI shows the scores achieved by the models when trained on the Short Jokes dataset. The DilBERT^2 was trained with the dynamic batching technique. The found accuracy and F1 scores are equal to 98% which matches the previously achieved result. The results on the ColBERT dataset are very similar, although the DilBERT^2 model was trained only on 1000 examples. The found accuracy and F1 score of DilBERT^2 are 96%, which is 2% less than the reported scores of the ColBERT model.

Method	Configuration	Accuracy	Precision	Recall	F1
DilBERTˆ2		0.983	0.985	0.980	0.983
TGtLL		0.986	0.986	0.986	0.986

TABLE VI: Comparison of DilBERT^2 against TGtLL reported performance on the Short Jokes dataset.

Model : Dataset	Configuration	Time	Accuracy	F1
DilBERT^2	Fixed Batching	1min 12s	0.957	0.96
DilBERT^2	Dynamic Batching	47s	0.962	0.960
ColBERT			0.982	0.982

TABLE VII: Comparison of performance on the task of humor detection on the ColBERT dataset.

Method	Configuration	Accuracy	Precision	Recall	F1
Decision Tree		0.560	0.560	0.560	0.560
SVM	sigmoid, gamma=1.0	0.543	0.542	0.543	0.541
Multilingual NB		0.580	0.569	0.580	0.570
XGBoost		0.616	0.602	0.611	0.591
DilBERTˆ2	Base-Cased: Fixed Batching	0.700	0.699	0.700	0.699
DilBERTˆ2	Base-Cased: Dynamic Batching	0.799	0.829	0.810	0.819
DilBERTˆ2	Large-Cased	0.649	0.649	0.649	0.646

TABLE VIII: Comparisons of different methods on the task of sentiment analysis on the grammatically correct version of the Dilbert dataset.

B. Sentiment Analysis

The comparison of the scores achieved by DilBERT^2 and other baselines on the sentiment analysis task is presented in Table VIII. The scores achieved for the DilBERT^2 models vary between 65% and 80%. The model still outperforms the other baselines by a big margin, however, it is not as reliable as in the humor detection task.

C. Dynamic Batching

Tables V and VII present the differences in training times and the found performance when training the DilBERT^2 models on datasets with around 3600 and 1000 examples accordingly. When training on the Pun Of The Day dataset, dynamic batching reduces the total number of padding tokens from 180,950 to 61,429 (66% less). That translates into training time that is 61% shorter than with the fixed approach. When training on the ColBERT dataset, the use of dynamic batching decreases the number of padding tokens from 50,600 to 20,328 (60% less). The model using the dynamic batching finished training 35% faster. In both cases, the use of dynamic batching does not indicate any negative effects on the final performance of the model.

VI. Responsible Research

Taking into account the ethical aspect of the conducted study is of utmost importance, especially in the domain of artificial intelligence (AI). For the language classification tasks, it is important to have varied and inclusive datasets to increase the chances of correct detection of text given by people from different backgrounds, using different sentence structures or vocabulary. In the case of creating a model in a specific domain, it is important to provide many examples of the domain-specific text, for the model to learn how to operate with commonly used sentence structure. Ethical and legal aspects enforce following strict scenarios in which the data can be used. To allow for comparison between different models and using this research as a baseline, there is a need to present the procedure of the conducted study in a detailed and objective way.

A. Research Integrity

It is necessary to state that the Dilbert and Garfield datasets do not serve the purpose of copying the artwork for financial gain or publicity, but are used for pure research applications. The data from the Dilbert dataset has been modified for the needs of the study. Special characters were removed for sentence clarity when scraping the data. In the data

pre-processing phase of training the model, two versions of the dataset were used. One was obtained from the scraper, the text in the other dataset was rewritten to be grammatically correct, with expanded contractions. Some of the text data were removed due to its strong correlation with the comic image. This however does not affect the ethical aspects of the dataset. What is important, is the inclusion of domain-specific data, in this case, comic language from software development offices, and adult life. When creating AI language models it is crucial to reason about ethical concerns of the developed project and reflect on its possible reactions to new input, provided by different people. Human language is complex and features like humor or sentiment can be expressed in a variety of ways in different daily life settings of different people.

B. Research Reproducibility

To provide a good starting point for future research, descriptions of the methodology and the experimental process were written with precise descriptions of decisions carrying consequences that influence the output of the models. As a result, it is expected that this research could be a better source of knowledge and provide insight into the task of humor and sentiment detection in comic texts.

VII. Discussion

DilBERT^2 for humor detection matches or outperforms the chosen baselines on different datasets. The difference the model achieves in the domain of comic text is significant when compared to basic classifiers. That indicates the need for using a complex NLP model when a linguistic expression like humor needs to be detected. When trained and tested on the grammatically correct dataset DilBERT^2 gives more accurate predictions than when the parsed version of the dataset is used. That finding confirms that the BERT model is capable of learning more complex linguistic features when it encodes more of its characteristics. Nevertheless, the scores achieved by DilBERT^2 on the comic-domain are significantly lower than the scores achieved on jokes datasets. The results of training DilBERT^2 on the portion of the ColBERT dataset suggest that the difficulty of correctly interpreting comic texts is greater than when the classification happens between jokes and news headlines, and that the lower score may not purely be a consequence of the small training dataset.

The adapted choices for hyperparameters and model functions for the training benefit the accuracy. The use of the dynamic batching technique indeed does not seem to affect the quality of predictions given by the model. Moreover, the benefits in time and energy savings provide a good incentive to use the technique instead of the fixed batching approach. In contrast with previous studies [6], the results achieved by training the models for a higher number of epochs did not improve the quality of the predictions given by the model.

The use of the pre-trained ColBERT model does not allow for equally reliable humor detection in comic texts. Hence, a conclusion could be drawn that the DilBERT^2 for humor detection is a reliable humor classifier for the domain of comic-text. Therefore, the model could be used to bring improvements to the concurrently developed comic-generating GANs by providing an auxiliary classification of the text included in the comics.

However, the proposed model does not perform as well on the sentiment analysis task. The performance measured on the comic-domain test set varied across multiple runs between 65% and 80%. Because of the multi-label nature of the classification problem, dataset expansion could lead to improved performance.

VIII. Conclusions and Future Work

This research has proven the effectiveness of using a fine-tuned BERT model for the task of humor detection on a new domain of comic texts. DilBERT^2 for humor detection, using the chosen parameters and training techniques achieves 80% accuracy on a comic-text dataset. The frequency of correct predictions could increase with a larger dataset of training examples, as the model is able to perform better on datasets that vary less in grammar and vocabulary but when the training is done on the same amount of examples. The potential benefit of using DilBERT^2 with the comic GANs could include affecting the background color or the positioning of the characters within the generated pictures, based on the humor of the classified text. Similar use of the DilBERT^2 for sentiment analysis requires improving its performance. That could also be potentially solved with more examples provided to the model. Results achieved by DilBERT^2 suggest that modern NLP models are able to perform difficult linguistic classification in varied domains, for which new applications could be found in the literature, music, or cinema.

References

[1] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[2] Chi Sun et al. “How to fine-tune BERT for text classification?” In: *China National Conference on Chinese Computational Linguistics*. Springer. 2019, pp. 194–206.

[3] Issa Annamoradnejad. “Colbert: Using bert sentence embedding for humor detection”. In: *arXiv preprint arXiv:2004.12765* (2020).

[4] Orion Weller and Kevin Seppi. “Humor detection: A transformer gets the last laugh”. In: *arXiv preprint arXiv:1909.00252* (2019).

[5] Michaël Benesty. *Divide Hugging Face Transformers training time by 2 or more*. June 2020. URL: https : / / towardsdatascience . com / divide - hugging - face - transformers - training - time - by - 2 - or - more - 21bf7129db9q-21bf7129db9e.

[6] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. “On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines”. In: *arXiv preprint arXiv:2006.04884* (2020).

[7] Tianyi Zhang et al. “Revisiting few-sample BERT fine-tuning”. In: *arXiv preprint arXiv:2006.05987* (2020).

[8] Wei Yang et al. “End-to-end open-domain question answering with bertserini”. In: *arXiv preprint arXiv:1902.01718* (2019).

[9] Kai Hakala and Sampo Pyysalo. “Biomedical named entity recognition with multilingual BERT”. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 2019, pp. 56–61.

[10] Yukun Zhu et al. “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*. 2015, pp. 19–27.

[11] Amil Merchant et al. “What Happens To BERT Embeddings During Fine-tuning?” In: *arXiv preprint arXiv:2004.14448* (2020).

[12] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv preprint arXiv:1706.03762* (2017).

[13] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[14] Pengcheng Yin et al. “Tabert: Pretraining for joint understanding of textual and tabular data”. In: *arXiv preprint arXiv:2005.08314* (2020).

[15] Louis Martin et al. “Camembert: a tasty french language model”. In: *arXiv preprint arXiv:1911.03894* (2019).

[16] Zhenzhong Lan et al. “Albert: A lite bert for self-supervised learning of language representations”. In: *arXiv preprint arXiv:1909.11942* (2019).

[17] Victor Raskin. *Semantic mechanisms of humor*. Vol. 24. Springer Science & Business Media, 2012.

[18] Issa Annamoradnejad and Zoghi Gohar. “ColBERT-Using-BERT-Sentence-Embedding-for-Humor-Detection”. In: *https://github.com/Moradnejad/ColBERT-Using-BERT-Sentence-Embedding-for-Humor-Detection* ().

[19] *Short Jokes Kaggle Dataset*. https://www.kaggle.com/abhinavmoudgil95/short-jokes.

[20] Diyi Yang et al. “Humor recognition and humor anchor extraction”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 2367–2376.