INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Humor and Offense Speech Classification and scoring using Natural Language Processing

Marcelo Custódio Mathias

Master in Data Science

Supervisors:

PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2022

Department of Quantitative Methods for Management and Economics
Department of Information Science and Technology

Humor and Offense Speech Classification and scoring using Natural Language Processing

Marcelo Custódio Mathias

Master in Data Science

Supervisors:

PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2022

Resumo

A identificação do humor e ofensa pode revelar-se uma tarefa árdua mesmo para os humanos. No entanto, é ainda mais desafiante traduzi-lo num processo lógico que uma máquina possa compreender.

Este trabalho pretende desenvolver modelos de aprendizagem automática que serão implementados para cumprir esta tarefa. Este estudo será baseado no workshop SemEval 2021, onde os participantes foram desafiados a detectar e classificar sentenças em relação ao humor e ofensividade, bem como detectar frases controversas (SemEval 2021 - Tarefa 7 - Detecção e Classificação de Humor e Ofensa), encorajando a utilização de estratégias algorítmicas de última geração focadas no processamento computacional da língua.

O objectivo é identificar e propor a melhor configuração para alcançar o melhor desempenho na Detecção de Humor e tarefas relacionadas, utilizando um conjunto de dados comum que agrega oito mil sentenças classificadas com os respectivos identificadores binário de humor e classificação, juntamente com os identificadores binários de controversas e classificação de ofensas.

Este documento apresenta uma solução para as tarefas apresentadas baseada no BERT (Bidirectional Encoder Representations from Transformers) que faz uso de Transformers, uma arquitetura de rede neuronais que permite interpretar as sentenças em ambos os sentidos (bidireccional), o que traz uma melhor percepção de contexto quando comparada com outras arquiteturas. Este estudo compara o desempenho de três variantes de BERT ($\text{BERT}_{\text{BASE}}$, DistillBERT, and RoBERTa), cada uma delas concebida para se adaptar melhor às diferentes tarefas utilizadas pela indústria e pelo meio académico. Concluiu-se que DistillBERT apresentou o melhor desempenho nas tarefas de Detecção de Humor e Classificação de Humor, enquanto RoBERTa foi mais preciso na tarefa de detecção de frases controversas. Finalmente, $\text{BERT}_{\text{BASE}}$ obteve a melhor performance na tarefa de Classificação de Ofensividade.

Palavras chave

Detecção de Humor, PNL, BERT, Detecção Controvérsia, Detecção Ofensa

Abstract

Identifying humor and offense may prove to be an arduous task even for humans. It is, however, even more challenging to translate it into a logical process that a machine can understand.

This work pretends to develop machine learning models which will be implemented to achieve this task. On this track, this study will be based on the SemEval 2021 workshop, where the participants were challenged to identify and score both humor and offense texts, as well as detect controversial sentences (SemEval 2021 - Task 7 - Detecting and Rating Humor and Offense), encouraging the use of current state-of-the-art algorithmic techniques in Natural Language Processing.

The objective is to identify and propose the most optimal setup to achieve the highest performance on Humor Detection and related tasks using a common dataset aggregating eight thousand sentences classified with their respective binary humor indicator and humor rating, along with binary controversial indicators and offense rating values.

This document presents a solution for the presented tasks based on BERT (Bidirectional Encoder Representations from Transformers) which makes use of Transformers interpreting the sentences in both directions (bidirectional), which brings a much higher context perception into the model. It will compare the performance of three different BERT variants ($\text{BERT}_{\text{BASE}}$, DistillBERT, and RoBERTa), each of them designed for better fit on different tasks used by industry and academia. Concluding that DistillBERT presented the most accurate results in the Humor Detection and Humor Rating tasks, while RoBERTa performed best in the controversial detection task. Finally, $\text{BERT}_{\text{BASE}}$ outperformed in the Offensiveness Ranking task.

Keywords

Humor Detection, NLP, BERT for Humor, Controversiality Detection, Offensiveness Detection

Acknowledgements

This document is the result of successful work developed during the last years at ISCTE - Instituto Universitário de Lisboa, which I am very grateful for providing me the opportunity to be part of this collaborative environment in which extensive scientific knowledge is a constant presence.

Then and most important, I need to thank my co-supervisors Professors PhD Fernando Batista and Phd Ricardo Ribeiro, who guided me through this challenge and supported me all the time. Their expertise in NLP is an inspiration that motivated me to try and experiment with different methods, which made the results presented here possible. My sincere thanks.

As well, I wish to thank the colleagues I met on the course and who were my partners during many different tasks, and were also present during all the challenging moments to motivate each of us to succeed.

Finally, I would like to thank my family, who made it all possible from the very beginning, when I decided to sign up for this master's course. Especially my wife Daniela who took responsibility for my tasks while I worked to achieve this dream. Thanks, this would never be possible without you, and this unquestionably also belongs to you.

My sincere thanks are extended to each and every one of you!

Lisboa, Outubro de 2022

Marcelo Mathias

Contents

List of Figures

List of Tables

1 Introduction

This chapter explains the purpose and motivation for this study by addressing the most recent development of Natural Language Processing, which embraces a scenario where human perceptions can be translated into machine algorithms, in this case, applied to Humor Detection. Based on that, it will propose the goals to be achieved, the research questions in which we will be focused on herein, and the methodology used. Finally, this chapter will briefly describe this document, aiming to explain its basic structure.

1.1 Motivation

> "*Humor is an experience that makes a person happy or amused. Throughout history, humans have been studying it from a psychological or linguistic perspective, but to see it through the eyes of a computer, which is basically figuring out the patterns and sequential repetitions in the textual content, is a challenging task for the field of NLP*" [7].

The hardness of comprehending humor is complemented by Subies et al. [8]:

> "*. . . when analyzing which parts of the text the models use for deciding whether the text is humorous or not are based on heuristics, such that whether or not the tweet represents a conversation. This shows that there is still much work to do until language models are able to understand the inherent semantics of the text so well that it can really understand the aspects of the texts, independently of the text form, that causes laughter. However, humor is an expression of high-level intelligence, expressed in sophisticated communication techniques, therefore only understanding the text meaning is probably not enough for many cases*".

Essentially, it means that the comprehension of humor is not some kind of action that a human can explain, or even write down a guideline demonstrating how it is done, in other words: it is not a predictable process. Therefore, translating it into a computer algorithm is equally difficult, and only recently, the scientific community has realized how to use NLP to make it possible with satisfactory performance.

NLP has seen some new models developed during the last decade that focus on large amounts of data in order to synthesize common NLP tasks. Models like GPT, GPT-2, and BERT are examples of tools that were developed to achieve those tasks in a way that their main effort is put into their principal construction phase, and their possible users improve it on a way that makes sense for their needs in an affordable and precise manner.

This study intends to focus on one of the most advanced used techniques nowadays based on BERT (Bidirectional Encoder Representations from Transformers) and on how the different implementations of BERT perform to execute the humor detection task. We will use a dataset provided by the workshop SemEval 2021 [9] analyzing different BERT variants comparatively over this data.

1.2 Goals

Humor detection is a highly complex and sophisticated human skill. Therefore, translating it into NLP models has been a challenge presented by other academic works. Although, Humor detection may not be the end goal, nowadays there are concerns not only centered on Humor itself but also on other related tasks such as the following: a positive-humored sentence can also be interpreted as offensive, or how funny is it? Thus, Humor detection is the shortest task currently being studied by academics; in fact, it is just the beginning of introducing new concepts related to Humor.

The main purpose of this work is to study it from an extended perspective, to track how the main BERT models are accurate enough not only to detect humor but also to support these other perspectives. To make it possible, this study will be based on the SemEval 2021 Workshop - Task 7 [9] which targets this more widely open perspective of Humor detection. The workshop was proposed in 2021 with some tasks aiming to extend from humor detection, to its controversiality, and offensiveness.

The workshop challenge was completed in 2021, and the proposal here is to compare this study with the workshop results, made available by the organizers after its conclusion. The proposed tasks are listed in the next sections.

1.2.1 Detecting and Rating Humor

First, the model must be able to identify whether or not a given sentence should be classified as humor and rank the sentences according to F1-score. Alongside this task, it is also proposed to rank its humorousness in a range from 1 to 5, which will be assessed using the RMSE (Root Mean Square Error).

1.2.2 Controversiality Detection

Observing the variance in humor, we can detect some disagreement on whether or not a sentence can be perceived as humorous, which indicates that not everyone agrees on whether that sentence is considered humorous, which suggests that it is controversial. In line with this concept, the workshop also asks the participants to classify whether or not the sentence is controversial. It is a binary task, in which the results are assessed using the F1-score.

1.2.3 Offensiveness

There are many types of humor, like the ones based on offensive content, which may be seen as funny by some people while being aggressive by others. That is a very common situation in Humor. The focus here is to rate how offensive the sentences are on a scale from 1 to 5, regardless of whether they are identified as humorous or not (even non-humored sentences can obviously be offensive). This is a regressive task assessed using the RMSE.

1.3 Research Questions

Based on the goals listed in the previous section, this study will aim to answer the proposed questions by applying the BERT and some of its variants. The questions will spread from Humor detection to other perspectives, as mentioned before. Other than the humor, this study will focus on identifying also the controversiality and offensiveness over the sentences, as synthesized below:

1. The main NLP industry used BERT variants models are successfully able to detect and rate humor, controversiality, and offensiveness?
2. The same BERT models may achieve high performance on different tasks (humor, controversiality, and offensiveness)?
3. What are the most suitable models for each of these tasks?

In response to these questions, we will experiment with the BERT models using the dataset provided by the SemEval 2021 - Task 7, then evaluate and compare the results, considering other workshop participants.

1.4 Methodology

The methodology chosen for this study is the CRISP-DM (Cross Industry Standard Process for Data Mining), which will be partially applied to this study scenario.

CRISP-DM is an industry-independent process model for data mining, as presented by Schröer et al. [10], and is comprised of six phases cyclically related to each other, which are the following: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These phases are applied to the project as described below:

1. Business Understanding: at this point, the project demands a comprehensive understanding of the data and proposed task. In a full-project scenario it would be a more investigative task to understand the expectations and the available data. In this study, the context is suppressed and synthesized by using a dataset that is constrained by business understanding.

2. Data Understanding: the majority of the data used in this study is clearly described in the workshop documentation, reducing the effort to understand the data, even though we do some additional exploration to realize the distribution of data according to the tasks we proposed. Again, in a full-project, the data collected from several sources would have to be explained by their respective data owners, and we would need to make sure that all available information would be correctly understood by the team before moving into the next phase. Since the dataset was built previously by the workshop, the data understanding is this case is disregarded.

3. Data Preparation: in order to make the given data work properly with BERT models, some adjustments are required to compile the sentences according to the supported format handled by these models. These adjustments will be described in Chapter 4.

4. Modeling: at this point, the possible models are selected according to the data available and the problem needs, which in this case will focus on BERT variants. It may be required to test different models and approaches to select the most adequate one. Also, the model setup must be arranged and executed respecting the desired goals.

5. Evaluation: the results and the process are reviewed and explained, and further actions may be taken. In this case, the results will be compared with the results shared by other workshop participants.

6. Deployment: the data analysis results are provided to the final users informing them whether the initial investigation concerns were successfully solved. Also, the final solution is deployed and made available to the final user, concerns about monitoring, maintenance, and learning must be considered here.

Since CRISP-DM is a cyclical process, it is possible to return to the first phase at any point and improve the overall solution. Likewise, when working with data, the team becomes more adept at understanding the concept and business as long as they validate partial results, which is very similar to the adaptive software development model, where a business partially implements small features over time, allowing it to achieve a complete and final solution.

1.5 Document Structure

This paper is composed of six chapters, briefly described here. Background: overviews the current state-of-the-art on NLP models which are referred to by other academic works with a focus on Humor Recognition. It then delves into BERT's basic architecture and variations, explaining their differences and main approaches for each. Literature Review: it describes the source base method to start this study, in a way that can be easily replicated in the future, which is connected to Related Work: where it presents similar humor studies focusing on related tasks and their solutions.

We will describe in the following chapter, Data, the dataset structure provided by SemEval 2021 Task 7. We intend to understand and analyze the information contained in it, giving us valuable information for evaluating the final result.

Afterward, the results are presented in the Experiments and Results chapter, wherein the main purpose is to compare the test results, and define the most effective approach to accomplish the proposed tasks. Additionally, we intend to compare the work herein built with the other workshop participants.

Finally, the Conclusion and Future Work Chapter summarizes the results and concludes the study by answering the research questions and considering future steps that could help improve it.

2 Background

This chapter intends to demonstrate the main concepts that support the BERT models, namely its architecture based on Encoders and Decoders, and Attention fundamentals. Then, it demonstrates the main specializations of BERT and the specific scenarios to which they are applied to.

Throughout recent literature, BERT models have been cited as being the state of the art for detecting humor, for this reason, this study will aim to build a comparative work between the main BERT variants.

2.1 Humor Detection in Text

The simplicity of the Humor detection studies can be attributed to when they were first published. From this perspective, what was state-of-the-art in 2005 is certainly not the case today.

It is worthwhile to look at those studies, specifically in a chronological review, as they help to clarify how the researchers and industry have evolved so far. Mihalcea and Strapparava [11] are a suitable example. It was published back in 2005, and their approach (Naïve Bayes with SVM text classifiers) may not be at the top anymore, but in 2005 it was pure innovation as the author notes:

> "*In this paper, we showed that automatic classification techniques can be successfully applied to the task of humor-recognition. Experimental results obtained on very large data sets showed that computational approaches can be efficiently used to distinguish between humorous and non-humorous texts, with significant improvements observed over apriori known baselines. To our knowledge, this is the first result of this kind reported in the literature, as we are not aware of any previous work investigating the interaction between humor and techniques for automatic classification.*"

On the other hand, nowadays we have authors working on novel additions to the Humor classification that may become the next state-of-art, like Sun et al. [12] which dealt with the dataset provided by a previous workshop task (SemEval 2020 - Humor in headlines news) in

order to suggest a more innovative approach to work with RoBERTa. The author reported: “The results prove that our proposed method can effectively improve the prediction effect of the neural network model.”

Similarly, Miraj and Aono [13] propose a framework (Integrating BERT and other Embeddings with Neural Network - IBEN) that combines different layers of BERT with a Bi-GRU neural network. The author affirms: “This framework performed very well on the task of humor detection”.

2.2 Text Representation

BERT is sourced from Transformer models proposed by Vaswani et al. [1]. The Transformer concept is basically an evolution of neural networks which makes better use of parallel processing to produce even more accurate results on NLP tasks, which will be better conceptualized in the following sections.

2.2.1 Transformer and Attention Architecture

The transformer is an evolution of LSTM cells where each one is composed of an Encoder and a Decoder based on Attention, and was originally created focused on language translation, targeting a better performance architecture (both on time and context understanding) as described by Vaswani et al. [1]: ”... a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.”.

The main Transformer architecture is demonstrated in Figure 2.1 where we can see the Encoder in the left lane, which is responsible for receiving the input and calculating a vector over each value (words in a sentence for example) that will represent the relationship of those inputs to each other. The main steps are described below:

1. Input Embedding: This is the first step, it is basically a normalization over space, for each input it is calculated a vector in space (considering the other inputs).

2. Positional Encoding: Then the encoder will apply a function that will represent the input according to its position in the sentence, meaning some context, adding it to the previous vector (data + context).

3. Multi-Head Attention: This layer is able to create a self-attention vector, which will consider each input (word) in relation to the others (relationship), then we have a vector with data + context + relationship. It is named multi-head, because the final result is an average of each input, to avoid the self-focus relationship being too strong.

4. Feed-Forward: It is basically a neural network layer that will walk through the vector, to assimilate learning. Here is where we see the training speed enhancement, the inputs are processed parallelly, meaning, better CPU usage.

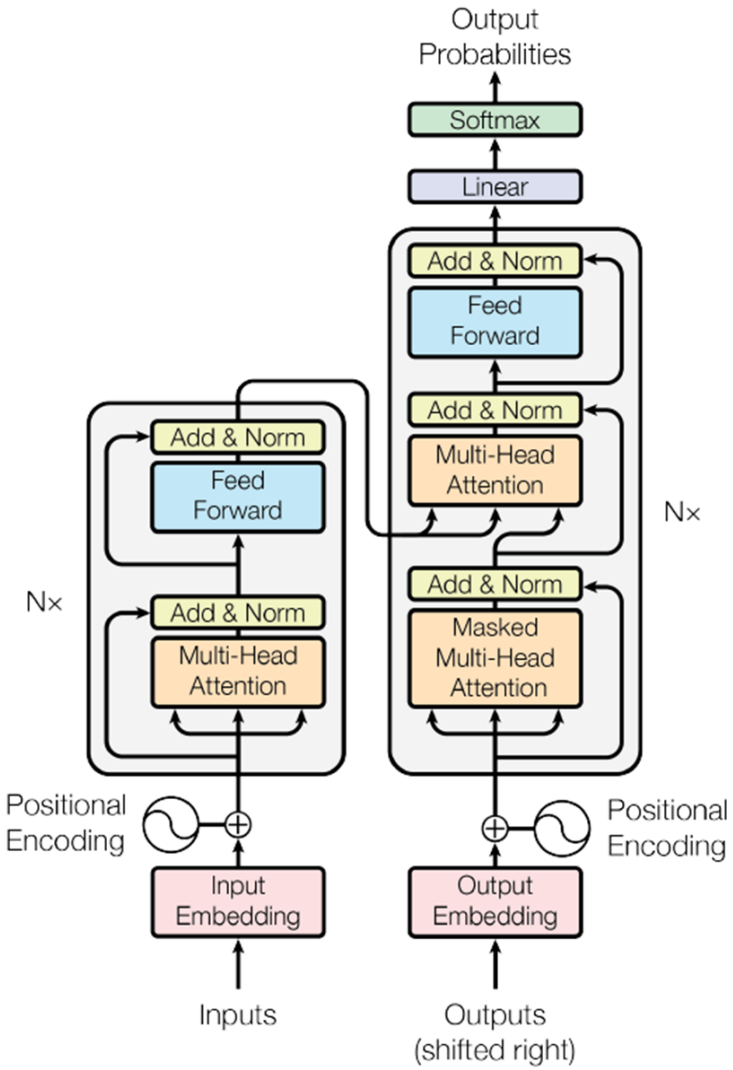


Figure 2.1: Transformer Architecture [1]

On the right side, we have the Decoder, which is responsible for taking the output variable and merging its vector with the input one, calculating the next word (it is originally focused on translating), according to the following steps:

1. Input Embedding: The same as the encoder, normalize the inputs in the space.
2. Positional Encoding: The same as encoder, will consider the position, then context.
3. Masked Multi-Head Attention: Similar to the encoder, but in the decoder, it does not

calculate all relationships between the inputs, only the previous one, masking the other ones to force it to learn to predict the next word.

4. Multi-Head Attention: Another layer of Attention, but now considering the input from Encode, that is where the relationship learning happens.

5. Feed Forward: Another neural layer, which will run in parallel, now with the input + output contexts.

6. Linear and Softmax: Normalization Layers.

7. Output: Next word in the sentence.

Each input is reprocessed by the decoder until the final word is reached. Consequently, some signal characters are needed to inform the model when to start or stop.

2.2.2 Evolution to BERT

The BERT makes use of the Encoder lane from Transformers to solve different tasks other than translation and works based on two distinct phases: the first is composed of Language Training, where the BERT is introduced to the base concepts of a specific language, and the result is usually defined as a pre-trained model. There are many BERT pre-trained models available from industry that are able to understand different languages targeting a large number of specific needs demanded by NLP researchers. The second phase is the fine-tuning approach, where the pre-trained model is taught to be able to handle a given task.

Additionally, BERT is an evolution of fine-tuning based on a pre-trained language model, where the key strategy relies on moving from unidirectional to bidirectional as explained in Devlin et al. [2]. Before BERT, the unidirectional technique was the most commonly used approach to fine-tuning pre-trained models, which at some point it presented limitations also reported by Devlin et al. [2]:

> "*The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-toright architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer. Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying finetuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.*"

2.2.2.1 Pre-train Phase

The pre-train phase was designed based on two unsupervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP).

The Masked Language Model is what gives BERT the bi-directional approach, as described by Devlin et al. [2] ”In order to train a deep bidirectional representation, we simply mask some percentage of the input tokens at random, and then predict those masked tokens.”, using the MLM implicitly carry over the context from the other tokens which make part of the sentence, thus making this phase bidirectional.

Along with MLM, BERT also applies a binary value striving to predict the next sentence (NSP - Next Sentence Prediction), learning from the relationship among the sentences. The following assumes that the author [2] used a set of real and random sequences: “Specifically, when choosing the sentences A and B for each pretraining example, 50% of the time B is the actual next sentence that follows A (labeled as IsNext), and 50% of the time it is a random sentence from the corpus (labeled as NotNext)”.

The joint of both tasks, MLM and NSP, makes BERT understand the main concepts and context of a token in a language, according to the source data where it is trained from. BERT used the BooksCorpus (800M words) and English Wikipedia (2,500M words), which were deployed as pre-trained models, from which the next phase, Fine-Tuning, could be carried out.

2.2.2.2 Fine-tuning

Fine-tuning phase is basically an additional train focused on an objective targeted task, and it works on the same architecture used for pre-train, wherein some adjustments are required as reported by Devlin et al. [2]:

> “*For each task, we simply plug in the task specific inputs and outputs into BERT and fine tune all the parameters end-to-end. At the input, sentence A and sentence B from pre-training are analogous to (1) sentence pairs in paraphrasing, (2) hypothesis-premise pairs in entailment, (3) question-passage pairs in question answering, and (4) a degenerate text-? pair in text classification or sequence tagging. At the output, the token representations are fed into an output layer for tokenlevel tasks, such as sequence tagging or question answering, and the [CLS] representation is fed into an output layer for classification, such as entailment or sentiment analysis.*”

This phase is relatively much faster than the previous one since the main effort is dedicated to pre-training the model. Usually, that is where industry starts working with BERT.

2.2.2.3 Main Architecture

Both phases run on the same architecture, by applying the MLM and NSP approaches, as demonstrated in Figure 2.2. And to make sure that the context is correctly understood, the input embedding is a key feature of this process. BERT bases the original embedding from the WordPiece embedding (with 30000 token vocabularies) [2] and in addition to the embedding, it sums the segment embedding and the position embedding, to make sure that both segment and position will be correctly carried into the model.

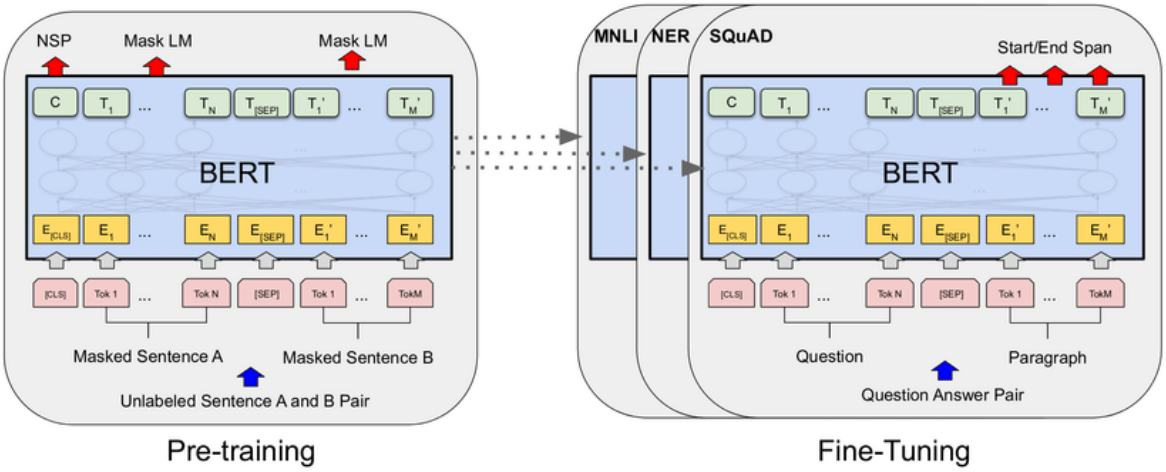


Figure 2.2: BERT Architecture [2]

It is also relevant to note that BERT makes use of special tokens to create a pair-sentence structure that takes the sequence into consideration. [CLS] is the token that represents the sequence start and [SEP] denotes the partition between both sentences, which creates segment sequence context to be learned by BERT [2].

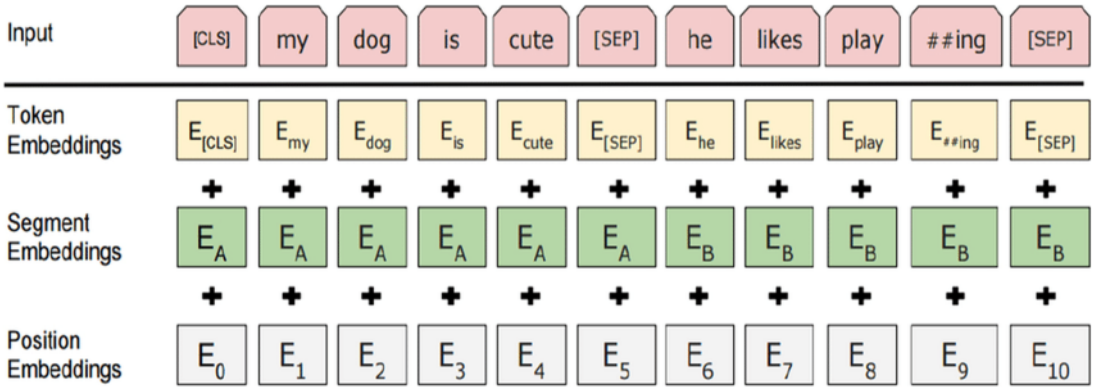


Figure 2.3: BERT Input Representation [2]

2.3 BERT Variants

When proposed, BERT was the only fine tuning-based representation model which could achieve state-of-the-art performance on large testing suites. Several variants were created over time, introducing various techniques over BERT to create numerous BERT-versions with specific specializations, some focused on improving time processing, others targeted some specific business domain, or improved quality, etc.

For the purpose of this study, we focus on three BERT versions: the original BERT, DistillBERT, and RoBERTa, aiming at the comparative results between them.

2.3.1 DistillBERT

While the Transfer Learning approach proposed by pre-trained NLP models has led to higher level performance rates, the cost of exponentially scaling these models may be too high, according to Sanh et al. [14]. Furthermore, finding models that fit on-device in real-time were the main purposes of Distillbert when it was proposed.

Distillbert is explained as a distilled version of BERT, which conceptually makes use of an advanced technique to reduce the complexity of a given model, whereas the efficiency remains very similar but the effort and cost are much reduced. The model was introduced in 2019 by Sanh et al. [14] making use of the knowledge distillation process as a compression technique.

The main idea behind model compression is to use a fast and compact model to approximate the function learned by a slower, larger, but a better performing model, by using the function learned on high-performance models to label pseudo data using a small neural net as reported by Bucilă et al. [15]. Additionally to this method, Hinton et al. [16] propose the distillation method which incorporates the use of a softmax-temperature function:

> "*In the simplest form of distillation, knowledge is transferred to the distilled model by training it on a transfer set and using a soft target distribution for each case in the transfer set that is produced by using the cumbersome model with a high temperature in its softmax. The same high temperature is used when training the distilled model, but after it has been trained it uses a temperature of 1.*"

The Knowledge Distillation process is a compression technique in which a compact model (the student) is trained to reproduce the behavior of a larger model (the teacher) or an ensemble of models [14].

When DistillBERT was proposed, Sanh et al. [14] demonstrated that the DistilBERT model retains 97% of BERT performance being significantly faster (around 61%) and 40% smaller.

2.3.2 RoBERTa

RoBERTa (Robustly Optimized BERT Approach) is an improved proposition based on the BERT architecture [17], after the authors identified that the original BERT was significantly undertrained. The updated study was based on a list of assumptions that, when compared, resulted in a list of better choices over the BERT pre-training process:

- More extensive training.
- Bigger batches of training over more data.
- Full sentences without Next Sentence Prediction (NSP).
- Dynamically change the masking pattern (MLM).
- Train over a new dataset (CC-News)

The robust model was comparatively studied with other models (like $\text{BERT}_{\text{LARGE}}$, $\text{XLNet}_{\text{LARGE}}$) on three different benchmarks: GLUE, SQuaD, and RACE. RoBERTa has demonstrated substantially improved performance on all comparisons, as reported by Liu et al. [17]: "Our improved pretraining procedure, which we call RoBERTa, achieves state-of-the-art results on GLUE, RACE and SQuAD, without multi-task finetuning for GLUE or additional data for SQuAD. These results illustrate the importance of these previously overlooked design decisions and suggest that BERT's pretraining objective remains competitive with recently proposed alternatives.".

3 Literature Review

This chapter will describe the research process for reviewing the literature on Humor Detection using NLP, targeting precedent studies provided by academic or institutional industry.

Humor Detection using NLP techniques seems to be a recent study area, due to the late evolution of text mining models, here we intend to walk through these works perceiving their chronological perspective and discern how Humor Detection has evolved so far.

Additionally, in this chapter it is also intended to identify and dig into the main technical concepts that will drive this study, looking for success and flaw results in order to find the most accurate process to execute Humor Detection.

To identify academic and institutional works that apply to the content proposed in this study, a 5-step process is employed, which simplifies the research work. The steps are described below:

1. Create a query for the past 6 years **[Exclusion Criteria]**, considering the current research understanding.
2. Filter the papers by title, abstract, and language (English or Portuguese), so the ones that cannot be identified with an objective relationship with the presented research questions are excluded. **[Exclusion Criteria]**
3. Initial content analysis comprehension.
4. Select the papers that are more significantly related to the current research, and also the ones with additional and useful information, disregarding the similar ones. **[Exclusion Criteria]**
5. Refine the query with new subjects that came up from the previous analysis, and repeat the steps, until reaching the more applicable articles and papers, which give the foundation concepts for the tools that make sense to this study.

In the literature review, we used two different data sources: Scopus and ACL Anthology. Both provide access to a large collection of documents with the possibility to easily replicate the queries reported in this chapter.

3.1 Humor Detection Studies Timeline

The first query objective is to have an overall perspective of how papers are distributed by country, language, and year. The query also is limited by the theme “Humor Classification” and by the subject areas “Computer Science” and “Mathematics”, since other areas (like Psychology) may cause huge noise in the results. The query was executed on the Scopus dataset with the following criteria: “TITLE-ABS-KEY (humor AND text AND classification) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MATH"))”, which provided a list of 30 documents.

This query is not focused on finding papers, rather the intention is to realize the overall scenario, which will indicate possible criteria for creating other queries that are more focused on our main subject.

The first criteria to be seen is how the papers are distributed by the years. As we can see in Figure 3.1, Humor classification keeps in an ascending trend, being the last 5 years when most of the work was produced since 70% of this query result was published between 2017 and 2021 (inclusive).

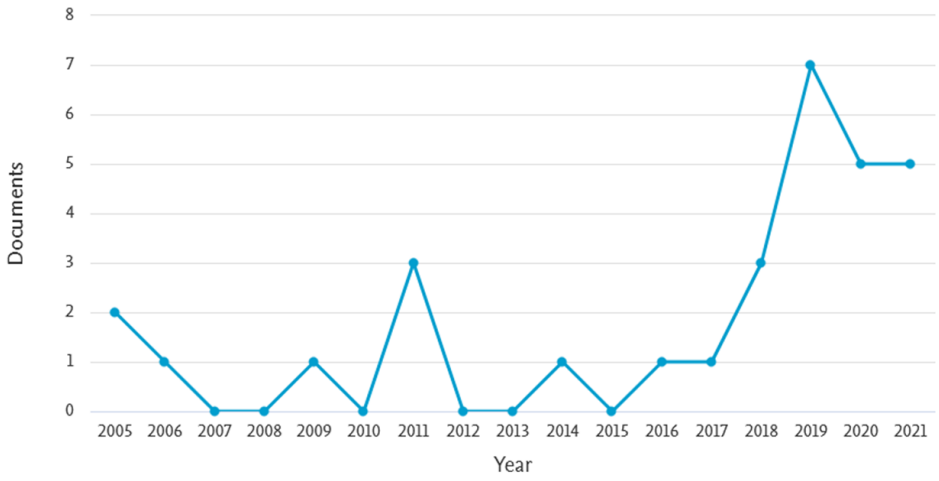


Figure 3.1: Literature Review - General Year Distribution

So, at this point, as we are looking for the current state-of-art in Humor detection, it makes sense to restrict the next queries to the following specific period: 2017-2021. But it also makes sense to look at what occurred in the past, as maybe there are some fundamental papers that may give us rich information.

3.2 Initial Work on Humor Detection

As described in the previous section, the query is now limited to only the early years (before 2017) and applies the exclusion criteria over the papers, in an effort to find the first

studies on Humor Detection. For this query, it is being used Scopus with the following filter: “TITLE-ABS-KEY (humor AND text AND classification) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MATH")) AND (LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2009) OR LIMIT-TO (PUBYEAR , 2006) OR LIMIT-TO (PUBYEAR , 2005)) ”.

Some of the papers that resulted in this query were more focused on the crowdsourcing process than models and their application. Therefore, we kept only the papers that were connected with humor detection, keeping in mind that these papers were discussing the very first attempts to pursue this task. However, they are helpful to understand how the researchers have developed on this topic.

It resulted in two documents after exclusion criterias:

- J. Costa, C. Silva, M. Antunes, and B. Ribeiro, “The importance of precision in humour classification,” Berlin, Heidelberg, pp. 271–278, 2011. [18]
- R. Mihalcea and C. Strapparava, “Making computers laugh: Investigations in automatic humor recognition - acl anthology,” 2005. [11]

3.3 Recent Work on Humor Detection

This query uses the same filter used in the first one but focused on the previous 5 years (2017-2021). The query is again executed on Scopus with the following filter: “TITLE-ABS-KEY (humor AND text AND classification) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MATH")) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017))”.

Using the output of this query, it is possible to see that most of the work being produced recently came from international workshops that encourage and promote NLP in several different languages. HAHA 2021 (part of IBERLaf 2021), which asked its attendees to solve the proposed tasks focused on the Spanish language; IberEval 2018, which focused on Iberian languages (Spanish, Portuguese, Catalan, Basque, and Galician); and SemEval which focused on more general tasks usually in English.

Despite our study being focused on English, it is still worth studying the strategies proposed for other languages. The strategies may be easily replicated, for this reason, papers on other languages are also being considered in this study.

Alternatively, this query raised one of the techniques which is the focus of this study: the BERT models and their variances (like Colbert and RoBERTa) as used by Annamoradnejad [19] and Liu et al. [17].

This query resulted in seven documents after exclusion criteria:

- A. Onan and M. A. Tocoglu, “Satire identification in turkish news articles based on ensemble of classifiers,” Turkish Journal of Electrical Engineering and Computer Sciences, vol. 28, pp. 1086–1106, 2020. [20]
- I. Annamoradnejad, “Colbert at haha 2021: Parallel neural networks for rating humor in spanish tweets,” 2021. [19]
- J. Mao and W. Liu, “A bert-based approach for automatic humor detection and scoring,” 2019. [21]
- J. Ortiz-Bejar, E. Tellez, M. Graff, D. Moctezuma, and S. Miranda-Jiménez, “Ingeotec at iberlef 2019 task haha,” 2019. [22]
- J. Ortiz-Bejar, V. Salgado, M. Graff, D. Moctezuma, S. Miranda-Jiménez, and E. S. Tellez, “Ingeotec at ibereval 2018 task haha: µtc and evomsa to detect and score humor in texts,” 2018. [23]
- Y. Kui, “Applying pre-trained model and fine-tune to conduct humor analysis on spanish tweets,” 2021. [24]

3.4 Deep diving into BERT studies

At this point, it makes sense to dive into BERT models being applied to humor detection. This query is still applied to the Scopus database, with the following filter: “TITLE-ABS-KEY (bert AND humor) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MATH"))”, which resulted in six documents after exclusion criteria:

- A. Ismailov, “Humor analysis based on human annotation challenge at iberlef 2019: First-place solution,” 2019. [25]
- G. G. Subies, D. B. Sánchez, and A. Vaca, “Bert and shap for humor analysis based on human annotation,” 2021. [8]
- K. Grover and T. Goel, “Haha@iberlef2021: Humor analysis using ensembles of simple transformers,” 2021. [26]
- P. Singh, A. Gupta, R. Sivanaiah, A. D. Suseelan, and M. Rajendram, “Techssn at haha @ iberlef 2021: Humor detection and funniness score prediction using deep learning techniques,” 2021. [7]
- R. Miraj and M. Aono, “Integrating extracted information from bert and multiple embedding methods with the deep neural network for humour detection,” International Journal on Natural Language Computing, vol. 10, no. 02, pp. 11–21, apr 2021. [13]
- Y. Sun, Y. Li, and T. Zhao, “The improved neural network model in humor detection with traditional humor theory,” pp. 549–554, 2021. [12]

3.5 Humor Detection from Other Sources

Alternatively to Scopus, and also to bring some comparison scenarios, here the plan is to use a different database which will be ACL Anthology. Unfortunately, the ACL does not provide the same filtering features as Scopus, so limiting its results is more difficult. Despite its limitations, the plan is to walk through the most relevant results and check what can be useful. Used filter: "https://aclanthology.org/search/?q=humor+dettection+text" ordered by Relevance.

The ACL Anthology does not give tools to limit the results, and restricting the query could cause it to remove papers that may be relevant for this study. In addition, duplicate results were found, making the original total of papers seem vaguely unreal. So, it was necessary to limit the query result to the first five pages, so the original query result (1040) was reduced to fifty papers. After the exclusion criterias, it resulted in seven documents:

- C. Zhang and H. Yamana, "Wuy at semeval-2020 task 7: Combining bert and naive bayes-svm for humor assessment in edited news headlines," 14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings, pp. 1071– 1076, 2020. [27]

- E. Simpson, E.-L. Do Dinh, T. Miller, and I. Gurevych, "Predicting humorousness and metaphor novelty with Gaussian process preference learning," Florence, Italy, pp. 5716–5728, Jul. 2019. [28]

- M. K. Hasan, W. Rahman, A. Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, Mohammed, and Hoque, "UR-FUNNY: A multimodal language dataset for understanding humor," EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 2046–2056, Apr. 2019. [29]

- P.-Y. Chen and V.-W. Soo, "Humor recognition using deep learning," New Orleans, Louisiana, pp. 113–117, Jun. 2018. [30]

- S. Kayalvizhi, D. Thenmozhi, and A. Chandrabose, "SSN_NLP at SemEval-2020 task 7: Detecting funniness level using traditional learning with sentence embeddings," pp. 865–870, Dec. 2020. [31]

- S. Mahurkar and R. Patil, "Lrg at semeval-2020 task 7: Assessing the ability of bert and derivative models to perform short-edits based humor grading," 14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings, pp. 858–864, May 2020. [32]

- V. Blinov, V. Bolotova-Baranova, and P. Braslavski, “Large dataset and language model fun-tuning for humor recognition,” pp. 4027–4032, Jul. 2019. [33]

3.6 Controversy and Offense Detection

To supplement literature relating to other proposed tasks, a different filter was applied to Scopus to access the content that is exclusively related to Controversial and Offensiveness detection. The filter in use is: “TITLE-ABS-KEY ((controversy OR offense) AND detection AND text) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MATH")) ”.

After applying exclusion criteria, we obtained six documents:

- A. Gupta, A. Pal, B. Khurana, L. Tyagi, and A. Modi, “Humor@iitk at semeval-2021 task 7: Large language models for quantifying humor and offensiveness,” Apr. 2021. [34]
- D. Thenmozhi, P. Nandhinee, S. Arunima, and S. Amlan, “Ssn_nlp at SemEval 2020 task 12: Offense target identification in social media using traditional and deep machine learning approaches,” Barcelona (online), pp. 2155–2160, Dec. 2020. [35]
- B. Huang and Y. Bai, “hub at SemEval-2021 task 7: Fusion of ALBERT and word frequency information detecting and rating humor and offense,” Online, pp. 1141–1145, Aug. 2021. [36]
- H. Al-Omari, I. AbedulNabi, and R. Duwairi, “DLJUST at SemEval-2021 task 7: Ha-hackathon: Linking humor and offense,” Online, pp. 1114–1119, Aug. 2021. [37]
- J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, “SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense,” Online, pp. 105–119, Aug. 2021. [38]
- K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocon, D. Puchalska, and P. Kazienko, “Controversy and conformity: from generalized to personalized aggressiveness detection,” Online, pp. 5915–5926, Aug. 2021. [39]
- R. Sivanaiah, A. D. S, S. M. Rajendram, M. Tt, A. P. Singh, A. Gupta, and A. Nanda, “TECHSSN at SemEval-2021 task 7: Humor and offense detection and classification using ColBERT embeddings,” Online, pp. 1185–1189, Aug. 2021. [40]
- S. Benslimane, J. Azé, S. Bringay, M. Servajean, and C. Mollevi, “Controversy detection: a text and graph neural network based approach,” Dec. 2021. [41]
- T. Raha, I. S. Upadhyay, R. Mamidi, and V. Varma, “IIITH at SemEval-2021 task 7: Leveraging transformer-based humourous and offensive text detection architectures using lexical and hurtlex features and task adaptive pretraining,” Online, pp. 1221–1225, Aug. 2021. [42]

3.7 Review Analysis

Based on the restricted papers, that Humor Detection has been under research since 2005. As reported by Mihalcea and Strapparava [11], in the hope of creating learning models that could synthesize the proper understanding of Humor, even though it is a somewhat challenging task that even a human cannot predict in some cases. Today, humor detection has become a largely explored topic in NLP research. It is becoming more feasible as the architecture needed to process this kind of problem is becoming available, and the latest presented models make it seem more possible.

Based on the background literature identified by this study, it is possible to define some different approaches and focus of research:

- Workshop tasks – Most of the work identified here was created to solve real tasks proposed by the workshop leaders, and so, as some papers are related to the same tasks, it is possible to compare them and identify the most performative approaches. Also the workshops are a valuable source of datasets for other studies, since a lot of effort is put into building up realistic data for the proposed tasks.

- Novel studies – Composed of papers that propose novel ways of handling the humor detection, these studies usually rely on the based background literature. They try to improve by walking on alternative paths. The proposal may be completely original or may be composed of a joint of other known approaches, and then the results reached by the proposed solution are compared with other known models when dealing with the same problem.

- New Datasets – Since it is difficult to get comprehensive data to work well with NLP tasks (particularly since the data needs to be identified), and because Humor classification is a very human-designed task, then the datasets require a great deal of manual labor, this leads to a complex effort to make it happen. Accordingly, some researchers work solely on developing functional approaches to creating trustworthy data and making them available to academics and industry to simplify and standardize studies.

3.8 Additional Related Work

In this chapter, we will discuss similar research that has been done in the area of humor detection, which has become a more realistic task, as architectures capable of solving this type of problem are now available, as well as the latest models recently created.

3.8.1 Humor Detection

As reported previously, the first academic paper on Humor Detection was published in 2005 by Mihalcea and Strapparava [11], where the authors used a Naive Bayes model with SVM text classifiers, and were able to achieve a relatively good result considering the available information at the time.

Later, in 2011, Costa et al. [18] proposed an SVM model improved by selecting a set of weak unlabeled examples from the testing set and reviewing them manually, considering them "weak". The sentences classified by the SVM under a doubtful area, thus improving the overall performance by using a semi-automated approach. With this strategy, the author reached a precision of 87.80% over the Jester Dataset, which contains 4.1 million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users, as reported by the author.

Recently, neural networks have reached the state-of-the-art in detecting humor in similar tasks, and many approaches to this technique have evolved the results. In 2016, Bertero and Fung [43] applied LSTMs in conjunction with Convolutional Neural Networks (CNNs) to predict humor (specific punchlines) in TV show dialogues and reported positive results: “We showed that our neural network is particularly effective in increasing the F-score to 62.9% over a Conditional Random Field baseline of 58.1%. We furthermore showed that the LSTM is more effective in obtaining a higher recall with fewer false positives compared to simple n-gram shifting context window features”.

Additionally, Chen and Lee [44] also proposed a CNN to predict the audience’s laughter, with better performance and optimistic results in a more realistic scenario. This was done using a dataset based on a corpus collected from TED talks with a higher volume of data.

Finally, Humor detection has been a challenging task proposed by several academic NLP workshops in order to challenge the participants to reach the best possible performance using available tools and propose new viable approaches. SemEval is one of those workshops, as cited previously, other workshops that may be noted are the following: the IberLEF, Field matters, WiNLP, and some others, with some of them being focused on specific areas, like the Clinical NLP Workshop. The IberLEF and SemEval annually propose specific tasks focused on Humor Detection., IberLEF names their task as the "HAHA task", these tasks are hardened every year to create more disruptive scenarios and explore other related tasks, like offensiveness and controversiality. The workshops are creative environments, from which the participants gain rich information about NLP by usually publishing papers discussing their approach. We will present and discuss some of them in the context of related work.

3.8.1.1 IberEval Task HaHa

IberEVAL was the previous name for the IberLEF workshop. The name was changed in 2019 and they presented a new task in 2018. We are no longer able to get a detailed description of this task, though we can infer the context by reading the produced papers. As described by Iberval 2018 [45] about the program: "...aims at encouraging and promoting the development of Human Language Technologies (HLT) for Iberian languages (Spanish, Portuguese, Catalan, Basque and Galician), by creating series of evaluation and a discussion forum about roberta-base (RobertaForSequenceClassification) Natural Language Processing systems on an ongoing basis".

The task is named Humor Analysis based on Human Annotation (HAHA) and it provides a series of Spanish tweets to be classified by their humor as described by Ortiz-Bejar et al. [23] (one of the participants):

> "*Humor Analysis based on Human Annotation (HAHA) asks for systems that classify tweets, in the Spanish language, as humorous or not. Also, it asks for systems that determine (rank) how funny the tweets are. Those two tasks are described by HAHA organizers as follows:*
>
> ***Humor detection:*** *determining if a tweet is a joke or not (intended humor by the author or not). The results of this task will be measured using F-measure for the humorous category and accuracy. F-measure is the primary measure for this task.*
>
> ***Funniness score prediction:*** *predicting a funniness score value (average stars) for a tweet in a 5-star ranking, supposing it is a joke. The results of this task will be measured using root-mean-squared error (RMSE).*"

In this case, the author worked with a multilingual sentiment analysis classifier proposed by them to handle the problem together alongside other known models: "Our approach consists of well-tuned μTC (SVM) and EvoMSA models to perform both classification and regression tasks". The authors demonstrated their results, but at this time, it is not yet possible to compare their approaches, since it was not possible to find other papers talking about the same task. And the task organizer does not provide a result list. But it is pertinent to note that the proposed solution by this author is based on EvoDAG, B4MSA (Twitter sentiment classifier), and FastText, which are not the focus of this study.

In 2019, a new HaHa task was proposed. For this task, it is possible to get three different solutions, again, the detailed information from the program is no longer available.

The task is similar to the previous year, as described by Ismailov [25]:

> "*The challenge asks to classify tweets in Spanish as humorous or not, and rate how funny they are on a scale from one (not humorous) to five. The dataset is a corpus of crowd-annotated Spanish-language tweets split into a train and a test*

set. The train set consists of 24000 tweets out of which 38.6% are considered humorous with an average rating of 2.05. The test set comprises 6000 tweets for which only text is given. There are two tasks: Humour detection and Funniness score prediction:

Humour Detection: the goal is to classify tweets into jokes (intended humor by the author) and not jokes. The performance is measured using F1 score.

Funniness Score Prediction: the goal is to predict a funniness score (average of crowd-sourced ranks) for a tweet supposing it is a joke. The performance is measured using root-mean-square error."

Mao and Liu [21] worked on a solution based on BERT, while Ortiz-Bejar et al. [22] solved the task by using μTC (SVM), and Ismailov [25] worked with BERT + naïve bayes classifier, being the last one the winner approach as described in Table 3.1.

Reference	Sub Task 1 - F1	Sub Task 2 -RMSE
Mao and Liu [21]	0.784	0.910
Ortiz-Bejar et al. [22]	0.788	0.822
Ismailov [25]	0.82	0.736

Table 3.1: IberLEF 2019 Task HaHa - Results

3.8.1.2 SemEval 2020 Task 7

According to the SemEval 2020 [46] program, the task was: to "estimate the funniness of news headlines that have been modified by humans using a micro-edit to make them funny" being a micro-edit explained by the following replacements cited by the author, as illustrated in Table 3.2.

Original Headlines	Substitute	Grade
Kushner to visit **Mexico** following latest Trump tirades	therapist	2.8
Hilllary Clinton Staffers Considered Campaign Slogan 'Because It's Her **Turn**'	fault	2.8
The Latest: BBC cuts **ties** with Myanmar TV station	pies	1.8
Oklahoma isn't **working**. Can anyone fix this failing American state?	okay	0.0
4 **soldiers** killed in Nagorno-Karabakh fighting: Officials	rabbits	0.0

Table 3.2: SemEval 2020 – Task 7 – Micro-Edit examples

The main task was split into two subtasks. In the first one, participants are asked to rate the humor (on a scale of 0-3) of an edited headline, and in the second, they are asked to pinpoint which version is the funniest given two different edited sentences from the same headline.

Kayalvizhi et al. [31] worked with several different models in order to support the task, Mahurkar and Patil [32] based their work on BERT and its variants (RoBERTa, DistilBERT and ALBERT). Finally, Zhang and Yamana [27] focused on BERT+EDA and BERT+NB-SVM, for the first and second tasks respectively. The comparative scores are displayed on the Table 3.3.

Reference	Sub Task 1 - RMSE	Sub Task 2 -Accuracy
Kayalvizhi et al. [31]	0.8447	0.5367
Mahurkar and Patil [32]	0.5331	0.6217
Zhang and Yamana [27]	0.5242	0.5331

Table 3.3: SemEval 2020 – Task 7 – Results

3.8.1.3 Humor Data Sources

Identifying datasets where we can apply Humor detection is not a basic task, since it requires a very specific and specialized process to be done manually. To improve this scenario some academic researchers are working to build richer and more trustful datasets, by dedicating exclusive resources to building those datasets. The following are two recently published datasets, in use by the industry:

- Large Dataset and Language Model Fun-Tuning for Humor Recognition (Russian Language) from Blinov et al. [33]. Which is composed of 300000 classified short texts in Russian. It was primarily created from another original data source, which had its size tripled using automated collection tools.

- UR-FUNNY: A Multimodal Language Dataset for Understanding Humor (English Language) from Hasan et al. [29]. This provides a diverse multimodal dataset, which gives a more sophisticated perspective on text extracted from TED videos.

For the tasks proposed in this study, and to be strictly constrained in a similar scenario used in the workshop evaluation, the models here proposed are not applied to these datasets.

3.8.2 Offensiveness Detection

In the last few years, the NLP community has been focusing on approaches to text classification other than Humor Detection. Offensiveness classification is one of the most studied approaches. Social networks have created a world where mass or individual communication can be done instantly, so it is difficult to predict when one of those communications may cause harm to a particular person or group. Sometimes causing really painful consequences. Based on this scenario, multiple studies have been applied to detect and therefore prevent offensive content in an affordable time to avoid hazardous consequences, as reported by Liu et al. [17]: "Anti-social online behaviors, including cyberbullying, trolling and offensive language, are attracting more attention on different social networks. The intervention of such behaviors should be taken at the earliest opportunity. Automatic offensive language detection using machine learning algorithms becomes one solution to identifying such hostility and has shown promising performance."

Different techniques are being validated by the scientific community. Ranasinghe and Hettiarachchi [4] presented a comparative work to accomplish the task "Multilingual Offensive Language Identification in Social Media" proposed in SemEval-2020. The main target of this study is to identify a better approach for a variety of languages (Arabic, Danish, English, Greek, and Turkish). The authors decided to compare three different machine learning architectures:

1. Convolutional Neural Network (CNN): the authors considered that a CNN could be adequate to identify offensiveness patterns in text: "Since offense is mostly a word pattern, we assumed that CNNs would be a suitable architecture to detect offensive sentences".

2. Recurrent Neural Network (RNN): these networks are designed to work with sequential data, and offensive text can be better identified when set in a specific sequential order within the sentence. That is the reason the authors decided to include this architecture, incorporating the following variants: LSTM, bi-directional LSTM, GRU, and bi-directional GRU.

3. Transformers (BERT): he authors also considered the BERT models, specifically the BERT multilingual variant and others specific to various languages. (AraBERT, DANISHBERT, XLNet-large-cased, GREEKBERT, and BERTURK).

Overall, the results proved that the Transformer BERT multilingual architecture could find slightly better results than CNN and RNN, but when using language specialized BERT models, the results were much better.

	Arabic	Danish	English	Greek	Turkish
CNN	0.70	0.72	0.79	0.80	0.67
RNN	0.66	0.71	0.77	0.71	0.66
BERT-mulilingual	0.77	**0.80**	0.82	0.77	0.70
BERT specialized	**0.79**	0.76	**0.85**	**0.81**	**0.79**

Table 3.4: F1 macro results in the comparative study for Multilingual Offensiveness Detection [4]

Another experiment made in 2019 by Liu et al. [5] compared a Logistic Regression model, an LSTM network, and a BERT model over the Offensive Language Identification Dataset (OLID) to achieve three different tasks:

1. Detect Offensives;
2. Identify offensiveness type (Insult, Threat or Unknown);
3. Identify the target (Individual, Group, Entity or Other).

The BERT model again presented better results for tasks A and C, but for task B it failed, as seen in Table 3.5.

	Identify Offensiveness	Offensiveness Type	Target
Logistic Regression	0.7102	**0.6028**	0.5607
LSTM	0.7166	0.5029	0.5056
BERT	**0.7826**	0.3830	**0.8435**

Table 3.5: F1 macro results in a comparative study for Offensiveness Detection [5]

3.8.3 Controversy Detection

The difference between controversy and offensive content is that controversy is not easily detectable even by humans, as reported by Benslimane et al. [41]:

> *"Controversial content refers to any content that attracts both positive and negative feedback. Its automatic identification, especially on social media, is a challenging task as it should be done on a large number of continuously evolving posts, covering a large variety of topics."*

As part of the early prediction of controverse sentences, Hessel and Lee [6] presented a study focused on learning transference across domains for controversy detection, which is vital since what may be controversial from one perspective can be neutral from another,

according to the author: “For example, we identify “break up” as a controversial concept in the relationships subreddit (a subreddit is a subcommunity hosted on the Reddit discussion site), but the same topic is associated with a lack of controversy in the AskWomen subreddit (where questions are posed for women to answer). Similarly, topics that are controversial in one community may simply not be discussed in another: our analysis identifies “crossfit”, a type of workout, as one of the most controversial concepts in the subreddit Fitness”. The main concept proposed in this work starts by applying the detection not only over the initial post (considering that it was tested on Reddit social network), but also in the comments, then the controversial detection could be better understood in that specific domain.

The author compared some different text models: HAND, TFIDF, W2V, ARORA, LSTM, BERT-LSTM, BERT-MP, and BERT-MP-512. Those models were applied over six different datasets extracted from diverse Reddit domain communities: AskMen (AM), AskWomen (AW), Fitness (FT), LifeProTips (LT), personal finance (PF), and relationships (RL). Again, BERT-related models were the ones that presented the best results: “In general, the best performing models are based on the BERT features, though HAND+W2V performs well, too.” Additionally the authors selected the BERT-MP-512 as the chosen model to apply in the experiment. Performance results are listed in Table 3.6.

	AM	AW	FT	LT	PF	RL
HAND	55.4	52.2	61.9	59.7	54.5	60.8
TFIDF	57.4	60.1	63.3	59.1	58.7	65.4
ARORA	58.6	62.0	60.5	59.4	57.2	62.1
W2V	60.7	62.1	63.1	61.4	59.9	64.3
LSTM	58.9	58.2	63.6	61.5	60.0	63.1
BERT-LSTM	**64.5**	**65.1**	**66.2**	65.0	65.1	67.8
BERT-MP	63.4	64.0	64.4	65.7	64.1	67.0
BERT-MP-512	63.9	64.0	64.7	**65.8**	**65.6**	67.7
HAND+W2V	61.3	62.3	64.9	63.2	60.0	66.3
HAND+BERTMP512	63.6	63.5	64.9	64.1	64.4	**68.0**

Table 3.6: Average accuracy in comparative study for controversiality detection provided by Hessel and Lee [6]

Considering the comparative work on this study, we can conclude that BERT is a good alternative to identify controversial sentences, even though the main target of this study was to check the domain learning transference which the author did not consider successful: “One promising avenue for future work is to examine higher-quality textual representations for conversation trees. While our mean-pooling method did produce high performance, the resulting classifiers do not transfer between domains effectively. Developing a more expressive algorithm (e.g., one that incorporates reply-structure relationships) could boost predictive performance, and enable textual features to be less brittle”.

4 Data

This chapter will present the data, and its characteristics, used by this study to achieve the main goal. Studying Humor Detection faces a common hazard: it is difficult to obtain datasets with sufficient data to learn from, especially when we try to classify Humor Range (how funny is a sentence?). For that reason, there are not too many datasets available to answer the questions proposed in this paper. Since we want to evaluate our proposal against the SemEval 2021 workshop [9], it makes sense that this study makes use of the dataset proposed by the workshop.

4.1 Data Collection

The workshop task provided three different datasets during the challenge execution, the train and dev datasets were public and available at the very first moment, the train is composed of 8000 sentences, and is therefore the source that should be applied to train the model. The dev dataset is composed of 1000 sentences to be used by the participants to evaluate their model locally. Finally, the workshop also provided the test dataset after ending the challenge. It was used by the organizers to classify the participant's models and rank them. It also has 1000 sentences.

As reported by workshop organizers [9], the data was sourced 80% from Twitter, and 20% from Kaggle Short Jokes dataset [47] by filtering specific targets (Sexism, Body Characteristics, Race, Sexual Orientation, Racism, Ideology, Religion, and Health) focusing also on keywords that would better represent hate speeches.

The author ensured that the dataset was properly balanced from a humor perspective using both traditional (setup + punchline or absurd context) and offensive methods. In order to keep a good source from Twitter, for humor data only declared US humorous accounts were selected, for the non-humorous, the accounts are more widely spread, but some targets were disregarded, like news sources, to avoid huge stylistic discrepancies.

To select offensive sentences the authors reported: "we identified the target, or butt, of the joke and made the assumption that a text could be potentially offensive to our annotators if the hate speech keyword was the target of the joke.", then the selection was made based on a hate speech list.

Finally, an online tool (Figure 4.1) was used to recruit annotators who were asked to answer the following questions about the selected sentences:

1. Is the intention of this text to be humorous? If positive, then it was asked to rate it (from 1-5)
2. Is this text generally offensive? If positive, then it was asked to rate it (from 1-5)
3. Is this text personally offensive? If positive, then it was asked to rate it (from 1-5)

The author complements: "For the humor rating, the user was also given the option to select 'I don't get it', meaning that they recognized by the structure or content that the text was intended to be humorous, but that they were unsure of why the text was funny. This is distinct from a rating of 1, which is a recognition of humor, with little appreciation for it".

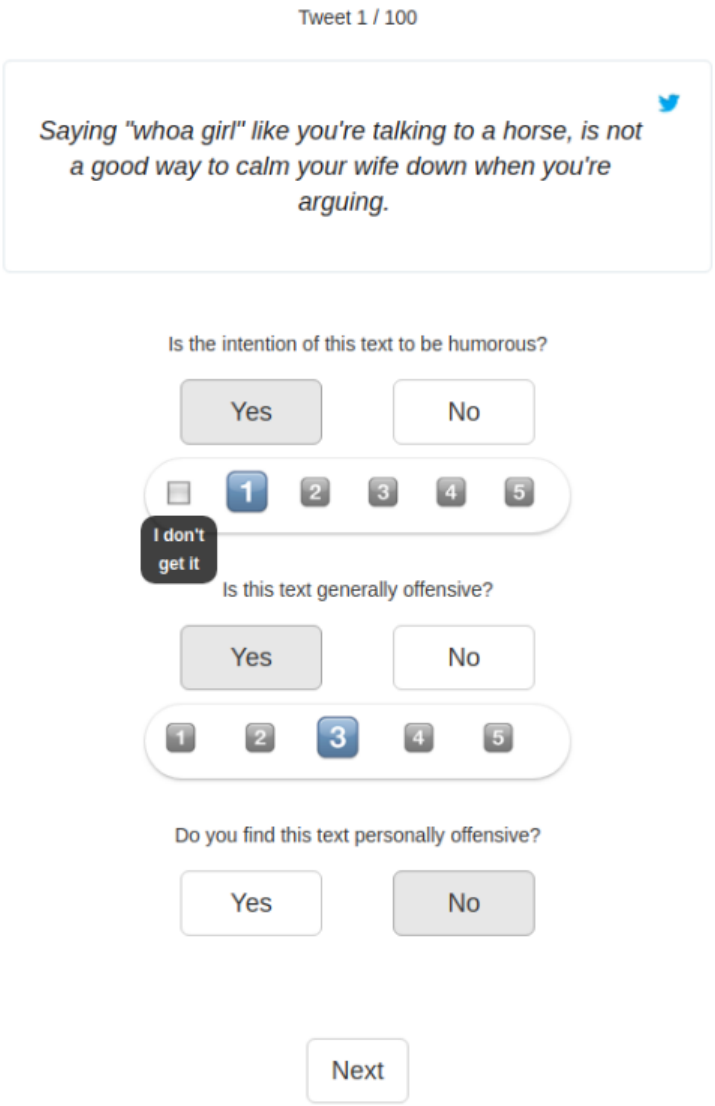


Figure 4.1: Example extracted from a tool used by annotators

4.2 Data Structure

The dataset is composed of an id, the sentence, and the classification variables, as shown in the Table 4.1.

id	text	is_humor	humor_rating	humor_controversy	offense_rating
1547	My son Luke loves that we named our children after Star Wars characters. My daughter Chewbacca not so much.	1	4.00	0	0.0
2194	When she asks you if she looks fat and you reply noo but it autocorrects to moo....	1	3.50	1	1.40
1296	What is a Cell? something you keep black people in	1	1.55	0	4.85
1903	Looking for black queer designers in NYC.	0	-	-	1.35
6409	Eliminate Anime and Islam, and you secure the the existence of western culture for eternity	0	-	-	2.4
4314	Eminem is afraid of giraffes. He doesn’t like their necks.	1	0.40	0	0.0
6	’Trabajo,’ the Spanish word for work, comes from the Latin term ’trepaliare,’ meaning torture.	0	-	-	0.0
112	The wise are not always quiet, but they know when to be	0	-	0	0.0

Table 4.1: Dataset Subset Example

The classification variables are defined as follow:

1. is_humor: A binary value (0/1) that indicates whether or not the sentence is classified as Humor. The distribution is 60% / 40% (humor / not humor), figure 4.2.
2. humor_rating: Classifies the humor level, from 1 to 5, only available when is_humor = 1, otherwise will be empty, figure 4.2.
3. humor_controversy: Binary value (0/1) which indicates whether or not the sentence is controversial, only available when is_humor = 1, otherwise will be empty. The distribution is 50% / 50% (controversial / not controversial), figure 4.3.
4. offense_rating: Classifies how offensive is a sentence, from 1 to 5, this variable is always supplied regardless of is_humor, since a sentence still can be offensive even not being funny, figure 4.3.

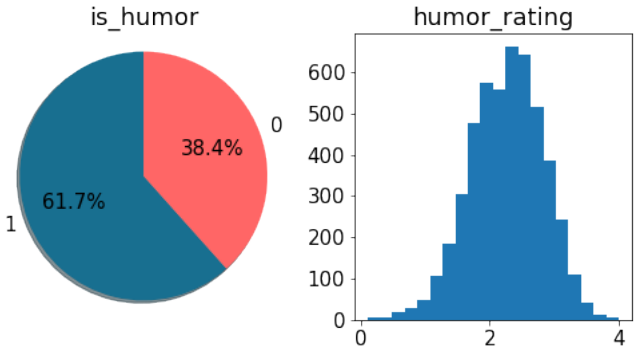


Figure 4.2: is_humor and humor_rating distribution

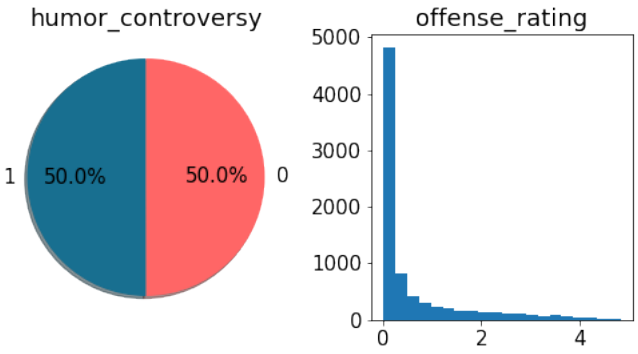


Figure 4.3: humor_controversy and offense_rating distribution

4.3 Data Preparation

The data provided by SemEval Task, as described on previously on this chapter, does not request any extra balance, since the authors considered it when building the data, but it still requires some setup according to the BERT requirements:

1. Special tokens “[CLS]” and “[SEP]”: to make BERT reach a better performance it is important to denote the start and end for each sequence, so every sentence must be set as “[CLS]” ... “[SEP]”.

2. Tokenize: The sentences must be translated into indexes that BERT will understand, we decided to use the BERTTokenizer provided by HuggingFace package.[1]

3. Padding: All input sentences must have the same size, it is needed to complement the smaller sentences to match the higher one with an empty value.

[1] https://huggingface.co/docs/tokenizers/index

4. Attention Mask: The BERT also requires an additional input which consists of an array with 0/1 marks, to distinguish what cells must be processed or ignored.

5 Experiments and Results

This chapter will analyze the results for each task and model executed comparatively, which from the CRISP-DM perspective it would be in the Evaluation phase. Additionally, we will compare the results with other participants' scores on SemEval 2021 Task 7 [9], to understand how successful the approach herein proposed is. To replicate the same workshop scenario, the most appropriate model will be chosen according to the dev dataset results, and only this one will be run against the test dataset.

5.1 Modeling

As seen in Chapter 3, the BERT and its variants are the widest machine learning algorithms used for humor detection. So to achieve the goals of this work it was decided to explore those algorithms and evaluate them against the provided dataset. Before starting to exploit them, we propose a baseline model that gave us the minimum score for the first task, and then, later on, we improve those models by using another architecture working with $\text{BERT}_{\text{BASE}}$, Distillbert, and RoBERTa pre-trained models. We work with Pythorch[1] and HuggingFace packages to handle the data and load pre-trained models respectively, running on Google Colabs GPUs.

5.1.1 Baseline

As described in Section 1.2 we are looking for more than humor detection: this work intends to also rank the humor, classify whether the sentence is controversial, and rank its offensiveness. Since all tasks are at some level similar, then the baseline was planned to be completed only for the first task, as it would be wasteful to execute a baseline for all tasks considering the time and cost to do so. The model was inspired by the tutorial proposed by Google [48], using a BERT-base-cased pre-trained model from HuggingFace package to calculate embeddings, and applying the features into an SVM model, running in an environment without GPUs, and applying the following steps:

1. *Data Prepare*

[1] https://pytorch.org/docs/stable/index.html

(a) Load ‘bert-based-cased’ tokenizer and model from HuggingFace package.

(b) Add special tokens [CLS] and [SEP].

(c) Tokenize sentences.

(d) Pad sentences.

(e) Create Attention Masks.

2. *Train and Evaluation*

(a) Create embeddings over train data: Calculate the features representing the embedding distribution for each sentence.

(b) Run SVM models

(c) Predict and score over dev data: Calculate F1 score

(d) Predict and score over test data: Calculate F1 score, to compare with the workshop result.

5.1.2 LLM Models

The second step is to validate each task with improved models running an LLM architecture on a GPU enabled environment. The initial version was based on the work proposed by Rebekah et al. [49] with the needed adjustments desired by each task. The data preparation execution is similar to the one used for baseline with the additional step to load the data into the Pytorch DataLoaders, which simplifies batch execution and GPU use, both on training and evaluation.

The steps are described below:

1. *Data Prepare*

(a) Load ‘bert-based-cased’ tokenizer and model from HuggingFace package.

(b) Add special tokens [CLS] and [SEP].

(c) Tokenize sentences.

(d) Pad sentences.

(e) Create Attention Masks.

2. *Train and Evaluation*

(a) Iterate over train dataloader: walk through the training sentences on GPU and update the model with back-propagation over 10 epochs.

(b) Set the model to evaluation.

(c) Predict and score over dev data: for binary tasks the metric is the F1-score, for regressive tasks is the RMSE.

(d) Predict and score over test data: repeat the eval on test data and compare the score with the workshop result.

5.2 Setup

Three BERT variants models were applied to each task: $BERT_{BASE}$, Distillbert, and RoBERTa, the models were fine-tuned over pre-trained versions provided by the HuggingFace package using their respective transformer versions (described below) with a classification layer on the top:

- bert-base-cased (BertForSequenceClassification): a BERT standard pre-trained transformer based on Devlin et al. [2], which differentiate cased/uncased words (case-sensitive).
- distilbert-base-uncased (DistilBertForSequenceClassification): an uncased distilled version of BERT based on Sanh et al. [14], which does not make a difference of cased words.
- roberta-base (RobertaForSequenceClassification): a case-sensitive robust version (RoBERTa) of BERT based on Liu et al. [17].

The experiments were executed on Google Colab environment with 1 active GPU running 10 epochs.

5.3 Results

After executing the models according to the previously explained process, we are able to evaluate the results for the tasks proposed by the workshop.

5.3.1 Humor Detection

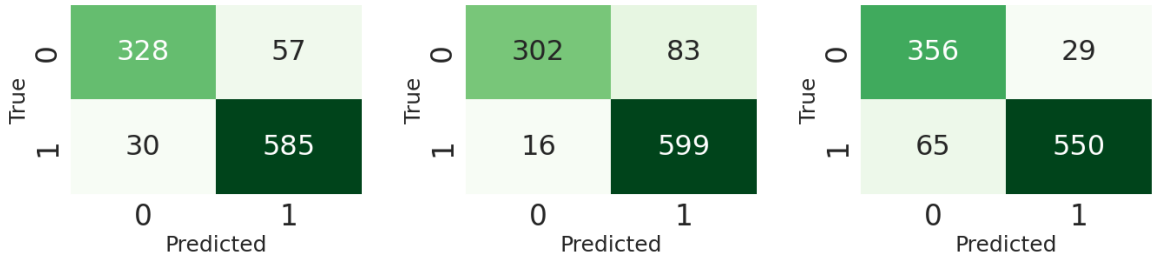
This task intends to identify whether or not the sentence should be classified as humor (binary task), for this task only, it was executed a baseline model to get an initial comparison, as described in Section 5.1.1. The results are listed in the Table 5.1.

Model	F1	Accuracy	Recall	Precision	Avg Runtime	Workshop Position
Baseline: BERT Embedding + SVM	0.9034	0.8780	0.9284	0.8798	06:32:15	40 of 44
$BERT_{BASE}$	0.9317	0.9160	0.9317	0.9317	00:03:08	34 of 44
DistillBERT	**0.9352**	0.9190	0.9512	0.9198	00:01:33	30 of 44
RoBERTa	0.9238	0.9080	0.9073	0.9409	00:03:06	34 of 44

Table 5.1: Humor Detection Task - Results - Test Data

Despite DistillBERT being the smallest and fastest model, it is the one that reached the best performance on this task (comparing the F1-score), even though, all studied BERT

models had very similar results. Additionally, by analyzing the confusion matrix in Figure 5.1, it is also possible to check that DistillBERT was the best to detect positive sentences, but the worst for negative ones (it is also perceived by the precision rate). The $\text{BERT}_{\text{BASE}}$ seems to be the most balanced model for this task, but since this task is restricted to the F1 score as proposed by the SemEval 2021 workshop, so the chosen for this task is DistillBERT.



$\text{BERT}_{\text{BASE}}$ DistillBERT RoBERTa

Figure 5.1: Confusion Matrix - BERT x DistillBERT x RoBERTa

5.3.2 Rating Humor

This task intends to classify how humorousness is a sentence (not too funny or too much funny) on a scale from 1 to 5 (regressive task), the comparative results are listed in the Table 5.2.

	RMSE	Avg Runtime	Workshop Position
$\text{BERT}_{\text{BASE}}$	0.6715	00:03:09	29 of 44
DistillBERT	**0.6694**	00:01:33	29 of 44
RoBERTa	0.6730	00:03:05	29 of 44

Table 5.2: Rating Humor - Results - Test Data

Again, the models presented very similar results (RMSE is slightly different between them) being DistillBERT the one which performed best in a much smaller timeframe. The workshop ranks this task by RMSE, the chosen model here will be DistillBERT.

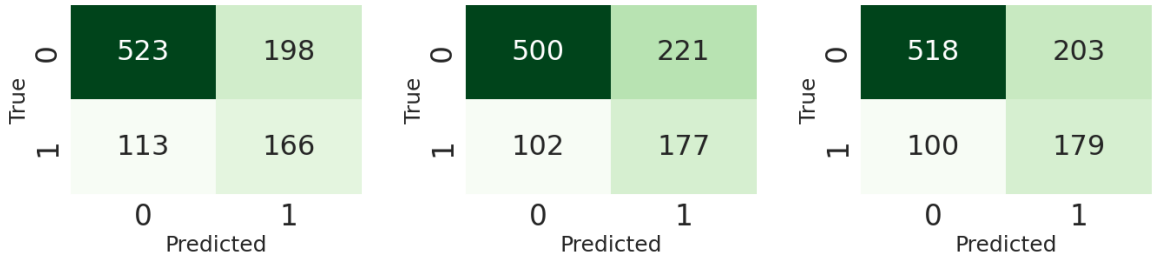
5.3.3 Controversiality Detection

This task proposes to identify whether or not a humorous sentence is controversial (when the humor is not easily perceivable). The comparative results are listed in the Table 5.3.

Model	F1	Accuracy	Recall	Precision	Avg Runtime	Workshop Position
$BERT_{BASE}$	0.5163	0.6890	0.5949	0.4560	00:03:08	21 of 44
DistillBERT	0.5228	0.6770	0.6344	0.4447	00:01:31	21 of 44
RoBERTa	**0.5416**	0.6970	0.6415	0.4685	00:03:11	20 of 44

Table 5.3: Controversiality Detection Task - Results - Test Data

Based on F1, RoBERTa was the model which better performed on this task, by checking the confusion matrix in Figure 5.2 it is possible to note that all models were really close to each other, it is also possible to note that none of the models reached a favorable result, also considering other participants' results, meaning that BERT probably is not the state-of-the-art for controversiality detection. The chosen model here is RoBERTa.



$BERT_{BASE}$ DistillBERT RoBERTa

Figure 5.2: Confusion Matrix - BERT x DistillBERT x RoBERTa

5.3.4 Rating Offensiveness

The intention here is to rate the sentence's offensiveness regardless of whether it is humorous or not. The comparative results are listed in the Table 5.4.

	RMSE	Avg Runtime	Workshop Position
$BERT_{BASE}$	**0.4872**	00:03:14	17 of 44
DistillBERT	0.4908	00:01:36	17 of 44
RoBERTa	0.5119	00:03:12	23 of 44

Table 5.4: Rating Offensiveness - Results - Test Data

In this task $BERT_{BASE}$ presented the best performance, even though it is not so far from DistillBERT, considering that it takes half of the time to be executed, and, in this RoBERTa has achieved the worst results, not even close to the other models. The workshop ranks this task by RMSE, and the chosen model here is $BERT_{BASE}$.

5.4 Comparative Analysis

After the conclusion of Humor Detection Task on SemEval 2021 workshop, the main participants and the organizers [38] deployed several papers discussing the best approaches which better supported each subtask. Each task had the players ranked by the respective indicator.

In this section, the intention is to compare and develop a critical analysis of other players' proposed solutions.

Additionally, for all tasks the organizers provided baseline results, trying to distinguish innovative ideas from just simple model implementation.

5.4.1 Humor Detection

The comparative results for Humor Detection are listed in Table 5.5

Team	F1	Chosen Model
PALI	0.9854	Unknown
stce	0.9797	Unknown
DeepBlueAI	0.9676	RoBERTa
SarcasmDet	0.9675	RoBERTa
baseline (BERT)	0.911	N/A
Current work	0.9352	DistillBERT

Table 5.5: Top results comparison - Detecting Humor Task

As reported by Meaney et al. [38] the teams PALI and stce did not publish a paper exposing their approach, so, no analysis can be done.

DeepBlueAI was the third in this task, and this team has built a comparative experiment using different training strategies [50]:

- Task-Adaptive Pre-training (TAPT): where the model is pre-trained in the provided Humor dataset;
- Pseudo-Labelling (PL): this strategy consists of using a labeling data model to predict the values on the test dataset, in order to increase the learning data. Some constraints are used, like using the prediction score of 0.8 (only sentences above this value were labeled as humor);
- Knowledge Distillation (KD): here the author uses intermediaries models which look for more objective results, known as hard target (humor/not humor) and soft target (probability), and finally add a loss reduction model;

- Adversarial Training (AT): adds noise to data, to improve the model robustness.

For this team specifically, the RoBERTa$_{\text{LARGE}}$ with TAPT, PL, and AT presented the best performance [50].

SarcasmDet [51] is also based on RoBERTa using different hyperparameters configurations, balancing the results by using a hard-voting ensemble technique, where the result is taken by the absolute majority response among all models.

5.4.2 Rating Humor

The comparative results for Humor Rating are listed in Table 5.6

Team	RMSE	Chosen Model
abcbpc	0.4959	ERNIE 2.0
mmmm	0.4977	Unknown
Humor@IITK	0.5210	DeBERTa
baseline (BERT)	0.800	N/A
Current work	0.6694	DistillBERT

Table 5.6: Top results comparison - Rating Humor Task

The team with higher performance “abcbpc” [3] worked with a comparative scenario between ERNIE 2.0 and DeBERTa models, using Multi Task Learning (MTL), where all tasks were based on the same model and the training from one task improves the robustness for other tasks. Additionally, the author also made use of ensemble, with cross-validation applying different hyperparameter settings, as described: “We adopt cross-validation for training as a way to improve the robustness of our model. We first divided the training set eight times by setting different random seeds. Therefore, 8 folds of data are generated, with 7000 training samples and 1000 validation samples in each fold. When fine-tune our model for each fold, the best model for each subtask at each fold of training is saved ... finally, we take the mean of all the best saved models after making predictions on the test set as the final results” as may be seen in Figure 5.3.

The ERNIE 2.0 model presented the highest performance.

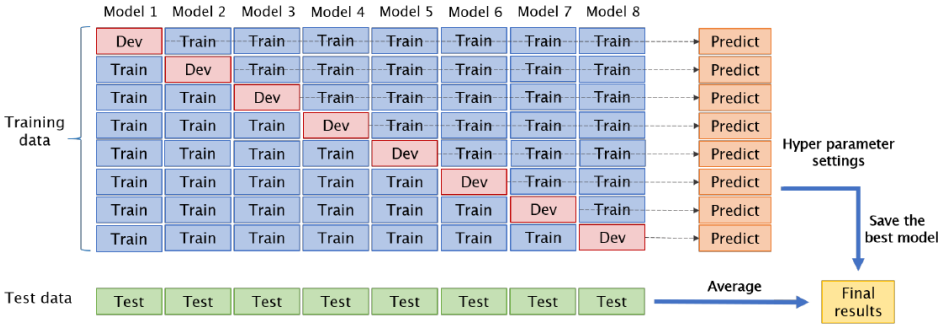


Figure 5.3: Training approach used by “abcbpc” team [3]

It is pertinent to note that ERNIE 2.0 is a multi-task learning model focused on capturing lexical, syntactic, and semantic aspects from text. In some cases, it outperforms BERT [52] and DeBERTa (Decoding-enhanced BERT with disentangled attention) is an evolution of RoBERTa models, which adds two novel techniques as reported by the author:

> “*The first is the disentangled attention mechanism, where each word is represented using two vectors that encode its content and position, respectively, and the attention weights among words are computed using disentangled matrices on their contents and relative positions, respectively. Second, an enhanced mask decoder is used to incorporate absolute positions in the decoding layer to predict the masked tokens in model pre-training.*” [53]

The team “mmmm” did not provide details about their solution.

Finally, the team “Humor@IITK” [54] presented a solution that made use of single and multi-task models like BERT, DistilBERT, RoBERTa, XLNet, Albert, Electra, DeBERTa, and ERNIE 2.0, by aggregating the weighted average output from best-performed ones, either from the single or multi-task approach.

5.4.3 Controversiality Detection

A comparison between the results achieved by the participants is listed in Table 5.7:

Team	F1	Chosen Model
PALI	0.6302	RoBERTa
mmmm	0.6279	Unknown
SarcasmDet	0.6270	RoBERTa / BERT
EndTimes	0.6261	Unknown
baseline (BERT)	0.6232	N/A
Current work	0.5416	RoBERTa

Table 5.7: Top results comparison - Controversy Detection Task

The controversy detection task does not present the same performance as seen in Humor detection, meaning that the research community has more space to propose better strategies to improve those results, as noted by Xie et al. [54]: "We conjecture this is because humor controversy is itself a highly subjective task, which is difficult even for humans". Currently, in this study, the experiment with RoBERTa underperformed below the task baseline. Probably due to the experiments being restricted to a single model, and not mixing different strategies as other teams made.

The team PALI which took the first place did not provide details about their approach, other than they used an ensemble RoBERTa large as described by Meaney et al. [38]. Also, the teams "mmmm" and "EndTimes" did not provide any details.

Analyzing the results from team "SarcasmDet" [51], as seen previously, they have used the method of hard-voting ensemble and hyperparameters tuning over RoBERTa-large, RoBERTa-base, BERT-large, and BERT-base models for the tasks "Humor Detection" and "Humor Rating", so the best results from previous tasks were applied on this task as well which acquired the third place.

5.4.4 Rating Offensiveness

Offensive rating task results are shown in Table 5.8.

Team	RMSE	Chosen Model
DeepBlueAI	0.4120	ALBERT / RoBERTa
mmmm	0.4190	Unknown
HumorHunter	0.4230	DeBERTa
abcbpc	0.4275	ERNIE 2.0
baseline (BERT)	0.5769	N/A
Current work	0.4872	BERT

Table 5.8: Top results comparison - Rating Offensiveness Task

For the offensive text rating, the winner team, DeepBlueAI [50], proposed a strategy where they stacked multiple pre-trained language models as reported by the authors: “We fine-tune two kinds of pretrained language models including ALBERT and RoBERTa with different training strategies such as pseudo labeling and knowledge distillation. Then, we stack them with a simple linear regression model. Experimental results show the effectiveness of this ensemble method and we win first place and third place for subtask 2 and 1a”.

The team HumorHunter [34] again worked with Multi Task Learning (MTL) based on DeBERTa leading them to the third place.

5.5 Summary

After the comparative experiments between BERT, DistillBERT and RoBERTa, it is possible to accomplish the following results:

- Humor detection: DistillBERT achieved the best performance concerning F1 (0.9352). In a fictional rank including the workshop results, it would rank at 30^{th} position (44 participants in total).
- Rating Humor: DistillBERT reached the best results with RMSE (0.6694), it would rank in 29^{th} position.
- Controversiality Detection: RoBERTa obtained the best results in this task concerning F1(0.5416), ranking in 20^{th} position. Even though it is better ranked at this task, this was the only one in which the baseline provided by the workshop was not reached. Eighteen teams could produce better results than compared to the baseline.
- Rating Offensiveness: $\text{BERT}_{\text{BASE}}$ acquired the best performance with RMSE (0.4872), it would rank in 17^{th} position.

6 Conclusions and Future Work

6.1 Conclusion

Humor detection and related tasks present a challenge for NLP models, firstly due to poor availability of training data, secondly, due to its subjectivity, it is hard to explain how a human being understands humor, although it is even more challenging when attempting to make sense of it in different cultures. Beyond that, the other tasks also bring a high level of singularity. Identifying controversial or offensive sentences is also highly dependent on who is consuming that information.

In this study we tried to reduce this doubtful environment, targeting on a specific academic task where different teams attempt to find the best approach, and also focusing on a specific tool: BERT. The workshop SemEval task is the proof which makes comparative work possible. However, it is pertinent to note that this approach does not capture all the subjectiveness that affects Humor and other tasks. Much more can be done on this subject.

To answer the first research question herein presented "The main NLP industry used BERT variants models are successfully able to detect and rate humor, controversiality, and offensiveness?" we may address the following topics:

- The answer to the humor detection question would be: Yes, BERT models have shown satisfactory results proving that they are the current state-of-the-art in NLP, for this task.
- In terms of the humor rating: Partially, the BERT-related models did not achieve perfect performance and even were outperformed by ERNIE 2.0 (based on the winner team results), although it is still quite close to the best results considering the tools currently available in the NLP community.
- When it comes to detecting controversy, the answer would be: Partially, BERT and its variants provided the best results, but it is still far from flawless. In addition, many teams could not even reach the organizer's baseline results, indicating this was the most difficult task to accomplish.
- Regarding Offensiveness rating: Again, the answer is partially, BERT and its branches

provided the most promising results, but the results are not necessarily close to what a human could execute.

For the second question “The same BERT models may achieve high performance on different tasks (humor, controversiality, and offensiveness)?” based on this document, the answer is No: since DistillBERT obtained the best results on Humor Detection and Rating, but it did not repeat similar results for other tasks, being RoBERTa the most indicated for Controversial Detection and $\text{BERT}_{\text{BASE}}$ for Offensiveness Rating. It did not demonstrate a clear pattern that could allow us to affirm that a specific model would fit perfectly for all different tasks.

Regarding the third question What are the most suitable models for each of these tasks?”, we point to the results in Section 5.5.

6.2 Future Work

After comparing this experiment’s results with other team members’ outputs, it is clear that working with different BERT variants into mixed solutions using Multi-Task Learning would be a very promising approach to achieve better results, and techniques that could augment the training data would also improve the model’s robustness. Also, other tools came out of this comparative work and could be used in a future implementation: Pseudo-labelling, Adversarial Training, Hard-Voting Ensemble, and Cross-Validation.

A number of BERT variants have shown promise in the Humor detection task, including DeBERTa and AlBERT.

Additionally, in the future, we can use additional Humor datasets to pre-train the model, before applying it to the main data.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," Oct. 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] C. Pang, X. Fan, W. Su, X. Chen, S. Wang, J. Liu, X. Ouyang, S. Feng, and Y. Sun, "abcbpc at SemEval-2021 task 7: ERNIE-based multi-task model for detecting and rating humor and offense," Online, pp. 286–289, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.35

[4] T. Ranasinghe and H. Hettiarachchi, "BRUMS at SemEval-2020 task 12: Transformer based multilingual offensive language identification in social media," Barcelona (online), pp. 1906–1915, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.semeval-1.251

[5] P. Liu, W. Li, and L. Zou, "NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers," Minneapolis, Minnesota, USA, pp. 87–91, Jun. 2019. [Online]. Available: https://aclanthology.org/S19-2011

[6] J. Hessel and L. Lee, "Something's brewing! early prediction of controversy-causing posts from discussion features," pp. 1648–1659, 2019. [Online]. Available: https://goo.gl/yHWeJp

[7] P. Singh, A. Gupta, R. Sivanaiah, A. D. Suseelan, and M. Rajendram, "Techssn at haha @ iberlef 2021: Humor detection and funniness score prediction using deep learning techniques," 2021.

[8] G. G. Subies, D. B. Sánchez, and A. Vaca, "Bert and shap for humor analysis based on human annotation," 2021.

[9] "Codalab - competition," 2021. [Online]. Available: https://competitions.codalab.org/competitions/27446

[10] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," vol. 181, 2021, pp. 526–534, cENTERIS

2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921002416

[11] R. Mihalcea and C. Strapparava, "Making computers laugh: Investigations in automatic humor recognition - acl anthology," 2005. [Online]. Available: https://aclanthology.org/H05-1067/

[12] Y. Sun, Y. Li, and T. Zhao, "The improved neural network model in humor detection with traditional humor theory," pp. 549–554, 2021.

[13] R. Miraj and M. Aono, "Integrating extracted information from bert and multiple embedding methods with the deep neural network for humour detection," *International Journal on Natural Language Computing*, vol. 10, no. 02, pp. 11–21, Apr. 2021. [Online]. Available: https://doi.org/10.5121%2Fijnlc.2021.10202

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[15] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," vol. 2006, 2006, pp. 535–541.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," Mar. 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[18] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, "The importance of precision in humour classification," Berlin, Heidelberg, pp. 271–278, 2011.

[19] I. Annamoradnejad, "Colbert at haha 2021: Parallel neural networks for rating humor in spanish tweets," 2021.

[20] A. Onan and M. A. Tocoglu, "Satire identification in turkish news articles based on ensemble of classifiers," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, pp. 1086–1106, 2020.

[21] J. Mao and W. Liu, "A bert-based approach for automatic humor detection and scoring," 2019.

[22] J. Ortiz-Bejar, E. Tellez, M. Graff, D. Moctezuma, and S. Miranda-Jiménez, "Ingeotec at iberlef 2019 task haha," 2019.

[23] J. Ortiz-Bejar, V. Salgado, M. Graff, D. Moctezuma, S. Miranda-Jiménez, and E. S. Tellez, "Ingeotec at ibereval 2018 task haha: µtc and evomsa to detect and score humor in texts," 2018.

[24] Y. Kui, "Applying pre-trained model and fine-tune to conduct humor analysis on spanish tweets," 2021.

[25] A. Ismailov, "Humor analysis based on human annotation challenge at iberlef 2019: First-place solution," 2019. [Online]. Available: https://www.kaggle.com/jhoward/

[26] K. Grover and T. Goel, "Haha@iberlef2021: Humor analysis using ensembles of simple transformers," 2021. [Online]. Available: https://www.fing.edu.uy/inco/grupos/pln/haha/

[27] C. Zhang and H. Yamana, "Wuy at semeval-2020 task 7: Combining bert and naive bayes-svm for humor assessment in edited news headlines," *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, pp. 1071–1076, 2020. [Online]. Available: https://aclanthology.org/2020.semeval-1.141

[28] E. Simpson, E.-L. Do Dinh, T. Miller, and I. Gurevych, "Predicting humorousness and metaphor novelty with Gaussian process preference learning," Florence, Italy, pp. 5716–5728, Jul. 2019. [Online]. Available: https://aclanthology.org/P19-1572

[29] M. K. Hasan, W. Rahman, A. Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, Mohammed, and Hoque, "UR-FUNNY: A multimodal language dataset for understanding humor," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 2046–2056, Apr. 2019. [Online]. Available: http://arxiv.org/abs/1904.06618http://dx.doi.org/10.18653/v1/D19-1211

[30] P.-Y. Chen and V.-W. Soo, "Humor recognition using deep learning," New Orleans, Louisiana, pp. 113–117, Jun. 2018. [Online]. Available: https://aclanthology.org/N18-2018

[31] S. Kayalvizhi, D. Thenmozhi, and A. Chandrabose, "SSN_NLP at SemEval-2020 task 7: Detecting funniness level using traditional learning with sentence embeddings," pp. 865–870, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.semeval-1.109

[32] S. Mahurkar and R. Patil, "Lrg at semeval-2020 task 7: Assessing the ability of bert and derivative models to perform short-edits based humor grading," *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, pp. 858–864, May 2020. [Online]. Available: https://arxiv.org/abs/2006.00607v1

[33] V. Blinov, V. Bolotova-Baranova, and P. Braslavski, “Large dataset and language model fun-tuning for humor recognition,” pp. 4027–4032, Jul. 2019. [Online]. Available: https://aclanthology.org/P19-1394

[34] A. Gupta, A. Pal, B. Khurana, L. Tyagi, and A. Modi, “Humor@iitk at semeval-2021 task 7: Large language models for quantifying humor and offensiveness,” Apr. 2021. [Online]. Available: http://arxiv.org/abs/2104.00933

[35] D. Thenmozhi, P. Nandhinee, S. Arunima, and S. Amlan, “Ssn_nlp at SemEval 2020 task 12: Offense target identification in social media using traditional and deep machine learning approaches,” Barcelona (online), pp. 2155–2160, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.semeval-1.286

[36] B. Huang and Y. Bai, “hub at SemEval-2021 task 7: Fusion of ALBERT and word frequency information detecting and rating humor and offense,” Online, pp. 1141–1145, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.160

[37] H. Al-Omari, I. AbedulNabi, and R. Duwairi, “DLJUST at SemEval-2021 task 7: Hahackathon: Linking humor and offense,” Online, pp. 1114–1119, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.155

[38] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, “SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense,” Online, pp. 105–119, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.9

[39] K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocon, D. Puchalska, and P. Kazienko, “Controversy and conformity: from generalized to personalized aggressiveness detection,” Online, pp. 5915–5926, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.acl-long.460

[40] R. Sivanaiah, A. D. S, S. M. Rajendram, M. Tt, A. P. Singh, A. Gupta, and A. Nanda, “TECHSSN at SemEval-2021 task 7: Humor and offense detection and classification using ColBERT embeddings,” Online, pp. 1185–1189, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.167

[41] S. Benslimane, J. Azé, S. Bringay, M. Servajean, and C. Mollevi, “Controversy detection: a text and graph neural network based approach,” Dec. 2021. [Online]. Available: http://arxiv.org/abs/2112.11445

[42] T. Raha, I. S. Upadhyay, R. Mamidi, and V. Varma, “IIITH at SemEval-2021 task 7: Leveraging transformer-based humourous and offensive text detection architectures using lexical and hurtlex features and task adaptive pretraining,” Online, pp. 1221–1225, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.173

[43] D. Bertero and P. Fung, "A long short-term memory framework for predicting humor in dialogues," San Diego, California, pp. 130–135, Jun. 2016. [Online]. Available: https://aclanthology.org/N16-1016

[44] L. Chen and C. M. Lee, "Predicting audience's laughter using convolutional neural network," Feb. 2017. [Online]. Available: http://arxiv.org/abs/1702.02584

[45] "Ibereval 2018," 2018. [Online]. Available: https://sites.google.com/view/ibereval-2018

[46] "Codalab - competition," 2020. [Online]. Available: https://competitions.codalab.org/competitions/20970

[47] M. Abhinav, "Kaggle short jokes dataset," 2017. [Online]. Available: https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes

[48] "Getting started with the built-in bert algorithm | ai platform training | google cloud." [Online]. Available: https://cloud.google.com/ai-platform/training/docs/algorithms/bert-start

[49] B. Rebekah, N. Luca, and V. Arnault-Quentin, "Github - rbkhb/nlp imc: Tutorial session on extracting information from social media data. part of the interacting minds center's nlp workshop at aarhus university on nov 7, 2019." 2019. [Online]. Available: https://github.com/rbkhb/NLP_IMC

[50] B. Song, C. Pan, S. Wang, and Z. Luo, "DeepBlueAI at SemEval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods," Online, pp. 1130–1134, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.158

[51] D. Faraj and M. Abdullah, "SarcasmDet at SemEval-2021 task 7: Detect humor and offensive based on demographic factors using RoBERTa pre-trained model," Online, pp. 527–533, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.64

[52] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.12412

[53] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," 2020. [Online]. Available: https://arxiv.org/abs/2006.03654

[54] Y. Xie, J. Li, and P. Pu, "HumorHunter at SemEval-2021 task 7: Humor and offense recognition with disentangled attention," Online, pp. 275–280, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.semeval-1.33

Detailed Results

A.1 Humor Detection

All results achieved during the Humor Detection task are described here.

A.1.1 $\text{BERT}_{\text{BASE}}$

Epoch	F1	Accuracy	Recall	Precision	Running Time
1	0,9227	0,9000	0,9446	0,9018	00:03:04
2	0,9170	0,8910	0,9525	0,8840	00:03:08
3	0,9230	0,9010	0,9383	0,9081	00:03:09
4	0,9247	0,9020	0,9525	0,8985	00:03:09
5	0,9213	0,8980	0,9446	0,8991	00:03:09
6	0,9220	0,8970	0,9636	0,8839	00:03:08
7	**0,9292**	**0,9090**	**0,9446**	**0,9142**	**00:03:08**
8	0,9182	0,8920	0,9589	0,8808	00:03:09
9	0,9106	0,8820	0,9509	0,8735	00:03:09
10	0,9227	0,8980	0,9636	0,8852	00:03:08

Table A.1: Detailed Results for BERT model on Humor Detection Task on Dev Dataset

After identifying the most optimal epoch for BERT on Humor Detection Task, it was run against the test dataset.

Epoch	F1	Accuracy	Recall	Precision	Running Time
7	0,9317	0,9160	0,9317	0,9317	00:03:08

Table A.2: Detailed Results for BERT model on Humor Detection Task on Test Dataset

A.1.2 DistillBERT

Epoch	F1	Accuracy	Recall	Precision	Running Time
1	0,9102	0,8820	0,9462	0,8768	00:01:33
2	0,9029	0,8780	0,8972	0,9087	00:01:33
3	0,9038	0,8720	0,9509	0,8610	00:01:33
4	0,9084	0,8790	0,9494	0,8708	00:01:33
5	0,9058	0,8790	0,9209	0,8913	00:01:33
6	0,9024	0,8760	0,9066	0,8981	00:01:33
7	0,9088	0,8790	0,9541	0,8676	00:01:33
8	**0,9132**	**0,8860**	**0,9494**	**0,8798**	00:**01:33**
9	0,9051	0,8710	0,9731	0,8459	00:01:33
10	0,9091	0,8800	0,9494	0,8721	00:01:33

Table A.3: Detailed Results for DistillBERT model on Humor Detection Task on Dev Dataset

After identifying the most optimal epoch for DistillBERT on Humor Detection Task, it was run against the test dataset.

Epoch	F1	Accuracy	Recall	Precision	Running Time
8	0,9352	0,9190	0,9512	0,9198	00:01:33

Table A.4: Detailed Results for DistillBERT model on Humor Detection Task on Test Dataset

A.1.3 RoBERTa

Epoch	F1	Accuracy	Recall	Precision	Running Time
1	0,7745	0,6320	1,0000	0,6320	00:03:05
2	0,8253	0,8010	0,7437	0,9270	00:03:05
3	0,8943	0,8620	0,9241	0,8665	00:03:06
4	0,8762	0,8530	0,8228	0,9369	00:03:05
5	0,9138	0,8900	0,9225	0,9053	00:03:05
6	0,9130	0,8880	0,9304	0,8963	00:03:05
7	0,9094	0,8790	0,9604	0,8634	00:03:05
8	**0,9196**	**0,8970**	**0,9320**	**0,9076**	**00:03:06**
9	0,8998	0,8760	0,8813	0,9191	00:03:05
10	0,8974	0,8690	0,9066	0,8884	00:03:05

Table A.5: Detailed Results for RoBERTa model on Humor Detection Task on Dev Dataset

After identifying the most optimal epoch for RoBERTa on Humor Detection Task, it was run against the test dataset.

Epoch	F1	Accuracy	Recall	Precision	Running Time
8	0,9238	0,9080	0,9073	0,9409	00:03:06

Table A.6: Detailed Results for RoBERTa model on Humor Detection Task on Test Dataset

A.2 Rating Humor

All results achieved during the Rating Humor task are described here.

A.2.1 $\text{BERT}_{\text{BASE}}$

Epoch	RMSE	Running Time
1	0,7245	00:03:08
2	0,7311	00:03:08
3	0,7318	00:03:08
4	0,7163	00:03:09
5	0,7163	00:03:08
6	0,7317	00:03:09
7	0,7130	00:03:09
8	0,7270	00:03:08
9	**0,7046**	**00:03:08**
10	0,7424	00:03:08

Table A.7: Detailed Results for BERT model on Rating Humor Task on Dev Dataset

After identifying the most optimal epoch for BERT on Rating Humor Task, it was run against the test dataset.

Epoch	RMSE	Running Time
9	0,6715	00:03:09

Table A.8: Detailed Results for BERT model on Rating Humor Task on Test Dataset

A.2.2 DistillBERT

Epoch	RMSE	Running Time
1	0,7503	00:01:33
2	0,7682	00:01:33
3	**0,7425**	**00:01:33**
4	0,7496	00:01:33
5	0,7683	00:01:33
6	0,7759	00:01:33
7	0,7676	00:01:33
8	0,7704	00:01:33
9	0,7718	00:01:33
10	0,7622	00:01:33

Table A.9: Detailed Results for DistillBERT model on Rating Humor Task on Dev Dataset

After identifying the most optimal epoch for DistillBERT on Rating Humor Task, it was run against the test dataset.

Epoch	RMSE	Running Time
3	0,6694	00:01:33

Table A.10: Detailed Results for DistillBERT model on Rating Humor Task on Test Dataset

A.2.3 RoBERTa

Epoch	RMSE	Running Time
1	0,7736	00:03:06
2	0,8064	00:03:05
3	0,7898	00:03:05
4	0,8008	00:03:05
5	0,8091	00:03:05
6	0,7558	00:03:05
7	0,7817	00:03:06
8	0,7696	00:03:05
9	**0,7459**	**00:03:05**
10	0,7808	00:03:05

Table A.11: Detailed Results for RoBERTa model on Rating Humor Task on Dev Dataset

After identifying the most optimal epoch for RoBERTa on Rating Humor Task, it was run against the test dataset.

Epoch	RMSE	Running Time
9	0,6730	00:03:05

Table A.12: Detailed Results for RoBERTa model on Rating Humor Task on Test Dataset

A.3 Controversiality Detection

All results achieved during the Controversiality Detection task are described here.

A.3.1 $\text{BERT}_{\text{BASE}}$

Epoch	F1	Accuracy	Recall	Precision	Running Time
1	0,5687	0,6830	0,6786	0,4895	00:03:09
2	**0,5876**	**0,6940**	**0,7078**	**0,5023**	**00:03:08**
3	0,5490	0,6960	0,6006	0,5055	00:03:09
4	0,5106	0,6990	0,5097	0,5114	00:03:09
5	0,4933	0,6960	0,4805	0,5068	00:03:08
6	0,4299	0,6950	0,3734	0,5066	00:03:09
7	0,4661	0,6770	0,4578	0,4747	00:03:09
8	0,5055	0,6870	0,5195	0,4923	00:03:09
9	0,4849	0,6920	0,4708	0,5000	00:03:09
10	0,4825	0,6890	0,4708	0,4949	00:03:08

Table A.13: Detailed Results for BERT model on Controversiality Detection Task on Dev Dataset

After identifying the most optimal epoch for BERT on Controversiality Detection Task, it was run against the test dataset.

Epoch	F1	Accuracy	Recall	Precision	Running Time
2	0,5163	0,6890	0,5949	0,4560	00:03:08

Table A.14: Detailed Results for BERT model on Controversiality Detection Task on Test Dataset

A.3.2 DistillBERT

Epoch	F1	Accuracy	Recall	Precision	Running Time
1	0,5567	0,7070	0,5974	0,5212	00:01:29
2	**0,5818**	**0,6880**	**0,7045**	**0,4954**	**00:01:31**
3	0,5169	0,6990	0,5227	0,5111	00:01:31
4	0,4764	0,7010	0,4416	0,5171	00:01:34
5	0,4738	0,7090	0,4253	0,5347	00:01:34
6	0,4862	0,7020	0,4578	0,5184	00:01:34
7	0,4612	0,7150	0,3961	0,5520	00:01:34
8	0,4135	0,7050	0,3377	0,5333	00:01:34
9	0,4648	0,7190	0,3961	0,5622	00:01:34
10	0,4850	0,7090	0,4448	0,5331	00:01:34

Table A.15: Detailed Results for DistillBERT model on Controversiality Detection Task on Dev Dataset

After identifying the most optimal epoch for DistillBERT on Controversiality Detection Task, it was run against the test dataset.

Epoch	F1	Accuracy	Recall	Precision	Running Time
2	0,5228	0,6770	0,6344	0,4447	00:01:31

Table A.16: Detailed Results for DistillBERT model on Controversiality Detection Task on Test Dataset

A.3.3 RoBERTa

Epoch	F1	Accuracy	Recall	Precision	Running Time
1	0,1944	0,6850	0,1234	0,4578	00:03:03
2	0,4155	0,6990	0,3474	0,5169	00:03:11
3	0,4800	0,7010	0,4481	0,5169	00:03:11
4	0,5394	0,6960	0,5779	0,5057	00:03:11
5	**0,5671**	**0,6840**	**0,6721**	**0,4905**	**00:03:11**
6	0,4472	0,6910	0,4058	0,4980	00:03:11
7	0,4735	0,7020	0,4351	0,5194	00:03:11
8	0,4000	0,6970	0,3279	0,5127	00:03:11
9	0,3811	0,6980	0,3019	0,5167	00:03:11
10	0,5116	0,6830	0,5390	0,4868	00:03:11

Table A.17: Detailed Results for RoBERTa model on Controversiality Detection Task on Dev Dataset

After identifying the most optimal epoch for RoBERTa on Controversiality Detection Task, it was run against the test dataset.

Epoch	F1	Accuracy	Recall	Precision	Running Time
5	0,5416	0,6970	0,6415	0,4685	00:03:11

Table A.18: Detailed Results for RoBERTa model on Controversiality Detection Task on Test Dataset

A.4 Rating Offensiveness

All results achieved during the Rating Humor task are described here.

A.4.1 $\text{BERT}_{\text{BASE}}$

Epoch	RMSE	Running Time
1	0,6457	00:03:11
2	0,6000	00:03:15
3	0,5911	00:03:15
4	0,6046	00:03:15
5	0,5926	00:03:14
6	0,6003	00:03:14
7	0,5937	00:03:14
8	0,5963	00:03:14
9	**0,5844**	**00:03:14**
10	0,5946	00:03:14

Table A.19: Detailed Results for BERT model on Rating Offensiveness Task on Dev Dataset

After identifying the most optimal epoch for BERT on Rating Offensiveness Task, it was run against the test dataset.

Epoch	RMSE	Running Time
9	0,4872	00:03:14

Table A.20: Detailed Results for BERT model on Rating Offensiveness Task on Test Dataset

A.4.2 DistillBERT

Epoch	RMSE	Running Time
1	0,6652	00:01:36
2	0,5850	00:01:36
3	0,5923	00:01:36
4	0,5993	00:01:36
5	0,5850	00:01:36
6	0,5997	00:01:36
7	0,5848	00:01:36
8	0,5829	00:01:36
9	0,5933	00:01:36
10	**0,5807**	**00:01:36**

Table A.21: Detailed Results for DistillBERT model on Rating Offensiveness Task on Dev Dataset

After identifying the most optimal epoch for DistillBERT on Rating Offensiveness Task, it was run against the test dataset.

Epoch	RMSE	Running Time
10	0,4908	00:01:36

Table A.22: Detailed Results for DistillBERT model on Rating Offensiveness Task on Test Dataset

A.4.3 RoBERTa

Epoch	RMSE	Running Time
1	0,7506	00:03:12
2	0,6613	00:03:12
3	0,6439	00:03:12
4	0,6267	00:03:12
5	0,6238	00:03:12
6	0,6196	00:03:12
7	0,6146	00:03:11
8	0,6059	00:03:12
9	0,6014	00:03:12
10	**0,5882**	00:**03:12**

Table A.23: Detailed Results for RoBERTa model on Rating Offensiveness Task on Dev Dataset

After identifying the most optimal epoch for RoBERTa on Rating Offensiveness Task, it was run against the test dataset.

Epoch	RMSE	Running Time
10	0,5119	00:03:12

Table A.24: Detailed Results for RoBERTa model on Rating Offensiveness Task on Test Dataset