BERT at SemEval-2020 Task 8: Using BERT to analyse meme emotions

Adithya Avvaru[1,2] and **Sanath Vobilisetty**[2]
[1] International Institute of Information Technology, Hyderabad, India
[2]Teradata India Pvt. Ltd, India
[1]adithya.avvaru@students.iiit.ac.in
[2]sanath.vobilisetty@teradata.com

Abstract

Sentiment analysis, being one of the most sought after research problems within Natural Language Processing (NLP) researchers. The range of problems being addressed by sentiment analysis is ever increasing. Till now, most of the research focuses on predicting sentiment, or sentiment categories like sarcasm, humor, offense and motivation on text data. But, there is very limited research that is focusing on predicting or analyzing the sentiment of internet memes. We try to address this problem as part of "Task 8 of SemEval 2020: Memotion Analysis" (Sharma et al., 2020). We have participated in all the three tasks of Memotion Analysis. Our system built using state-of-the-art pre-trained Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) performed better compared to baseline models for the two tasks A and C and performed close to the baseline model for task B. In this paper, we present the data used for training, data cleaning and preparation steps, the fine-tuning process of BERT based model and finally predict the sentiment or sentiment categories. We found that the sequence models like Long Short Term Memory(LSTM) (Hochreiter and Schmidhuber, 1997) and its variants performed below par in predicting the sentiments. We also performed a comparative analysis with other Transformer based models like DistilBERT (Sanh et al., 2019) and XLNet (Yang et al., 2019).

1 Introduction

Social Media (Mandiberg, 2012) has gained a lot of traction since it's inception. Almost 50% of the world's population is on social media. Be it Facebook, Instagram, Twitter or any other social media platform, content generation is growing exponentially day by day. Recently, there has been a surge of using memes[1] as a communication medium in social media platforms. Memes can take many forms; the most prominent usage of memes consisted of images combined with text. Memes, being images, can have text in one or more languages. This makes it even more difficult for an algorithm to understand and decode its sentiment or any other characteristic. Of late, smaller videos as memes are also gaining popularity. Irrespective of the type of meme, the meme may get changed, remixed or recreated while communicating through social media networks (French, 2017). Memes are used in contexts involving political discussion (Nave et al., 2018), to add a sarcastic perspective to the discussion, for social purposes, etc.

Meticulously analysing the memes can help us understand the involvement of societal factors like race and gender (Milner, 2013), implications on culture and the values promoted by the memes. Moreover, analysing the meme's underlying emotion can help us understand and possibly eradicate propagation of fake news, offensive content through memes (earlier identification of offensive content is only based on text (Zampieri et al., 2019)) and might also help in the prevention of malicious content. Image memes are also used to understand race and gender discourse on social media platforms, 4chan and Reddit (Milner, 2013). Memes, in future, might become an integral part of most of the people. Understanding the meme emotions will help us understand societal transformation over years or even decades.

[1] https://en.wikipedia.org/wiki/Internet_meme

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1094–1099
Barcelona, Spain (Online), December 12, 2020.

Meme ID	Meme Description
1	Consider a situation where in a kid standing on an already broken windshield of a car and the father saying ”THAT’S: MY DAUGHTER”
2	Consider a picture containing black and white persons and the black person saying ”RACISM EXISTS AMONG ALL RACES IN THE WORLD... WHITE PEOPLE ARE JUST BETTER AT IT LIKE MOST THINGS.”
3	Describing the effect of one of the Government schemes to curb black money, PM Modi trolled black money holders saying ”SONS AND DAUGHTERS WHO PUTS THEIR PARENTS IN OLD AGE HOME NOW PUTTING 2.5L IN THEIR ACCOUNTS TODAY WON’T I GET BLESSINGS OF THOSE PARENTS?”

Table 1: Sample memes description

Considering the role of internet memes in a wide variety of aspects of life, we participated in “SemEval Task 8: Memotion Analysis” (Shifman, 2019) to contribute to the development of research on meme emotion analysis. The task contains three individual tasks, as described in Section 3. We build a text-based model BERT (Devlin et al., 2018) and later apply fine-tuning to fit our train data. We are largely successful in producing better results in SubTask A, where we achieved Macro-F1 score of 0.3323 compared to baseline score of 0.2176 and slightly better results in SubTask C, where we achieved Macro-F1 score of 0.3038 compared to baseline score of 0.3009. Our model for SubTask B with Macro-F1 score of 0.4942 performed close to baseline Macro-F1 score of 0.5002.

2 Related Work

Recently, both Natural Language Processing and Vision & Image Processing research communities have been looking memes as a source of potential research. The remainder of this section talks about works related to meme analysis. The type of meme used in the communication is directly correlated to the nature of the topic in the social media and it is demonstrated that memes highlight the semantic context of the discussions on social media platforms (French, 2017). Memes are used in analysing serious social issues such as homophobic bullying of lesbian, gay, bisexual, transgender and queer (LGBTQ) community and establish collective identity (Gal et al., 2016).

Another area of focus is to understand the reasons for the spread of memes (or any content). Some of the reasons are novelty, simplicity, coherence and proselytism (a condition in which the meme provokes the other users to spread it further) (Chielens and Heylighen, 2005; Nave et al., 2018). Five other reasons for meme participation in social media are spread, emotional attachment to the users, ability to spread through different channels of communication, ability to add new meanings to existing image or text and provocation of other users to spread the meme (Milner, 2016).

Though there is a lot of research happening around internet memes, there is very little attention towards analysing meme emotions. Hu and Flaxman (2018) built a multi-modal deep neural network architecture to infer the emotional state of the user and predict the emotion word tags attached by the users to their Tumblr posts. However, this work does not focus on identifying the sentiment or variants of the sentiment like humour, sarcasm and motivation. This is our target area of research in this paper.

The structure of the paper is as follows. In Section 3, we give an overview of the tasks and dataset used for the experiment. Section 4 describes the experimental setup which includes data pre-processing, model description and training strategies, while Section 5 discusses the experimented results of various models. Finally, we conclude with concluding remarks and future direction of research in Section 6.

3 Tasks and Dataset Description

The data provided by the organizers has 6601 and 1878 samples for train data and test data respectively. The data need some cleansing process as training text is missing in some samples, labels are missing in some other. Class-wise number of samples in the training data for all the tasks after cleaning are shown in the Table 2. The cleaning process is described in Section 4.1.

3.1 Task A

Given an internet meme, the task is to determine the sentiment - positive, neutral or negative (3-class classification task). For instance, the memes 1 and 2 of the Table 1 have positive sentiment and meme 3 has a very negative sentiment (treated in this task as negative sentiment).

3.2 Task B

In this task, we have to identify the type of humour. Here, humour is categorised into four types - sarcasm, humour, offensive and motivational. Hence, this task is organized into 4 classification sub-tasks: sarcastic vs non-sarcastic, humorous vs non-humorous, offensive vs non-offensive and motivational vs non-motivational. As per the training data, the memes 1 and 3 in the Table 1 are humorous unlike meme 2. Similarly, the memes 2 and 3 are twisted sarcastic and the meme 1 is general sarcastic (both types are considered as sarcastic for this task). The memes 2 and 3 are offensive unlike the other one. Finally, the memes 1 and 2 are non-motivational unlike meme 3 which is motivational.

3.3 Task C

Contrary to Task B in Section 3.2, this task evaluates the extent to which each individual humour (defined in Task B) is being expressed. Like Task B, there are 4 sub-task here as well. The sarcasm detection here is not sarcastic vs non-sarcastic but a 4-class classification problem dealing with non-sarcastic, general sarcastic, twisted meaning and very-twisted meaning. Similarly, humour is further classified into non-funny, funny, very funny and hilarious. Offensive memes are further categorised into not offensive, slight offensive, very offensive and hateful offensive. Motivation sub-task here is the same as that of Motivation sub-task in Task B as both are 2-class classification problems. Though both the memes 1 and 2 of the Table 1 are categorised as sarcastic memes in Task B, Task C classifies meme 1 as a slight sarcastic meme and meme 2 as a very-twisted sarcastic meme.

Task	**Type of Problem**	**Labels for Classification**		**# Train Samples**
Task A	Sentiment Classification	Positive		3864
		Neutral		2070
		Negative		576
Task B, Task C	Sarcasm Detection	Non-sarcastic		1445
		Sarcasm	General	3239
			Twisted	1447
			Very Twisted	363
	Humour Detection	Non-Humorous		1545
		Humour	Funny	2282
			Very funny	2063
			Hilarious	603
	Offense Detection	Not offensive		2536
		Offense	Slight	2398
			Very	1360
			Hateful	207
	Motivation Detection	Motivational		4206
		Non-motivational		2277

Table 2: No of samples per classification label for the tasks after cleaning the data. The labels ‘General’, ‘Twisted’ and ‘Very Twisted’ are combined to form a class ‘Sarcasm’ for Task B. All are separate classes for Task C. Same is the case with ‘Humour’ and ‘Offense’.

4 Experimental Setup

4.1 Data Preparation

We employed the following steps for cleaning both the train and test data:

1. Removed all the empty samples present in the train data set.
2. Punctuation marks like exclamation (!) are used to express surprises, emotions, excitement etc in English text. Hence, we decided not to remove punctuation marks. However, we replaced consecutive instances of the same punctuation mark with only one instance of it.
3. As both the train and test data are Optical Character Recognition (OCR) extracted from the meme images, the data contains watermarks, some background texts, random website details, etc., which are removed during the cleaning process.

4. We have identified contracted words (for example, *we've*, *won't've*, etc,.) and replaced them with their corresponding English equivalents (in this case, *we have*, *will not have*, etc,.).

5. Simple spell correction like removing repetitive characters in the word. For instance, "soooooo niceeee" is converted to "so nice". This also might help in the reduction of feature space.

6. Other pre-processing steps include removing the URLs and @mentions. However, we decided to include hashtags as hashtags play an important role in sentiment evaluation. For example, the addition of the hashtag *#poor* changed the sentiment of the sentence *Made $174 this month, I'm gonna buy a yacht!* from slightly positive to negative.

4.2 Model Description

After data is pre-processed, we now have the cleaned text which is ready for training. There are two different approaches to training the data - (1) training from scratch with random initial model parameters and (2) applying the technique of transfer learning by fine-tuning the already trained (on large datasets) models. Since text type of data are predominantly sequences, we conducted experiments with models - Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), Bidirectional LSTM(BiLSTM) (Zhang et al., 2015), Stacked LSTM and Convolution Neural Networks - LSTM(CNN-LSTM) stacked model. We also conducted experiments with state-of-the-art Transformer based models like BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019) and XLNet (Yang et al., 2019). The BERT model (when trained using HuggingFace[2](Wolf et al., 2019)) need data in a certain format concerning separators and class labels and we followed the below steps to prepare the data into BERT compatible format and finally fine-tune the model:

1. Tokenize the cleansed text data

2. Create attention masks based on the padding done (as the sentences are of different lengths)

3. Fine-tune the pre-trained BERT model so that the model parameters will conform to the input training data

Similarly, HuggingFace (Wolf et al., 2019) version of XLNet needs BERT-like fine-tuning steps to fine-tune XLNet model. However, DistilBert model is trained using finetune[3] which provides fine-tuning APIs for NLP tasks, whose APIs signature is inspired by scikit-learn[4].

4.3 Training

Before working on test data, we split the cleaned train data into train and validation sets to conduct initial experiments. Using the new train data and validation data, we trained the models on LSTM variants (described in Section 4.2) and calculated the precision, recall and Macro-F1 scores on the validation data. The initial idea of choosing these models is because of their better performances on sequences data. However, the results are not encouraging. Then we started looking into fine-tuning transformer-based models - BERT(bert-base-uncased), DistilBERT (internally uses bert-base-uncased English version of BERT) and XLNet (xlnet-base-cased). We addressed the following problems that are encountered during training the models that are described in Section 4.2:

- **Over-fitting problem**: All the training models have a huge number of parameters which led to over-fitting of the models. We incorporated Dropout layers and early stopping to avoid over-fitting.

- **Class Imbalance problem**: Considering the data in Table 2, it is easily understood that the data is hugely imbalanced. For example, in Task A, the number of positive, neutral and negative samples are 3864, 2070 and 576 respectively. Similarly, class imbalance is present for other tasks and sub-tasks. When we train on this data, the generated model is skewed towards the majority class and hence the prediction performance is poor, specifically for the samples from minority class. We applied the over-sampling technique in all the tasks to address the class imbalance problem.

[2] https://huggingface.co/

[3] https://finetune.indico.io/

[4] https://scikit-learn.org/

5 Results

We experimented with LSTM variants for all the tasks on validation data. As the results are not encouraging even after oversampling, we quickly switched to start-of-the-art Transformer based models (mentioned in Section 4.2) for training the data in all the tasks and sub-tasks. These models are trained on full train data and prediction results are obtained for test data provided by the organizers. The results of Transformer-based models for all the tasks are shown in the table 3.

A quick analysis from the Table 3 show that our BERT model outperformed other Transformer-based models DistilBERT and XLNet in the majority of tasks and sub-tasks except both Motivational Sub-tasks of Tasks B and C. The prime reason to experiment with DistilBERT after conducting experiments with BERT is its application of Knowledge distillation concept (Bucila et al., 2006; Hinton et al., 2015) to generate a compact and compressed model while preserving larger part of BERT functionality. Sanh (2019) showed that DistilBERT is 40% smaller in size, 60% faster than BERT while retaining 97% of BERT's language understanding capabilities. However, DistilBERT performed slightly better (highlighted in the Table 3), only for motivation sub-tasks, probably because motivation sub-task is a two-class problem.

Model		**DistilBERT**		**BERT**		**XLNet**	
Task	**Sub-task**	**Individual Sub-task F1-Score**	**Average Task F1-Score**	**Individual Sub-task F1-Score**	**Average Task F1-Score**	**Individual Sub-task F1-Score**	**Average Task F1-Score**
Task A	**Sentiment**	0.3027	0.3027	0.3323	0.3323	0.2592	0.2592
Task B	**Humour**	0.4711	0.4660	0.4878	0.4942	0.3211	0.3612
	Sarcasm	0.4012		0.4432		0.3513	
	Offensive	0.4579		0.5123		0.3712	
	Motivation	**0.5339**		0.5334		0.4010	
Task C	**Humour**	0.2446	0.3063	0.2342	0.3038	0.1721	0.2312
	Sarcasm	0.2441		0.2355		0.1672	
	Offensive	0.2025		0.2121		0.1846	
	Motivation	**0.5339**		0.5334		0.4010	

Table 3: Results for all the tasks and subtasks

We also performed experiments with XLNet, a generalized bidirectional autoregressive model, which outperforms BERT in 20 NLP tasks and is designed to overcome the limitations of BERT model. Despite its huge success compared to BERT, XLNet models failed to perform in this task probably because of its complexity and lack of availability of large training dataset. Based on our experimental results, we stood at ranks 22, 16 and 16 for tasks A, B and C respectively. The score for tasks B and C are averaged F1-scores of their corresponding sub-tasks.

6 Conclusion and Future Work

Meme analysis is not on NLP researchers' radar couple of years ago; however, it is gaining importance, thanks to the advancements in internet and faster evolution of internet memes. We also tried to understand the sentiment and sentiment categories of memes by participating in "SemEval 2020 Task 8". We built models which performed reasonably well and outperformed baseline models in two tasks. We would like to focus on visual sentiment analysis (Hu and Flaxman, 2018) along with text sentiment analysis which is very popular and has a wide range of applications in areas like code-mixed sentiment analysis, irony detection, hate speech analysis, offence classification, emotion analysis, evaluating rumours, detecting sarcasm, humour or even motivation and many more. Visual sentiment analysis is a little more complex task as the sentiment message is embedded in different layers of image abstraction. As we have meme images, text and corresponding labels, we would like to extend this work and build models (possibly multi-modal models) for combined sentiment analysis across different aspects of humour, sarcasm and motivation.

References

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos, editors, *KDD*, pages 535–541. ACM.

Klaas Chielens and Francis Heylighen. 2005. Operationalization of meme selection criteria: Methodologies to empirically test memetic predictions. Citeseer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805.*

Jean H. French. 2017. Image-based memes as sentiment predictors. *2017 International Conference on Information Society (i-Society)*, pages 80–85.

Noam Gal, Limor Shifman, and Zohar Kampf. 2016. “it gets better”: Internet memes and the construction of collective identity. *New media & society*, 18(8):1698–1714.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop.*

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

Anthony Hu and Seth Flaxman. 2018. Multimodal Sentiment Analysis To Explore the Structure ofEmotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 350–358.

Michael Mandiberg. 2012. *The social media reader.* Nyu Press.

Ryan M Milner. 2013. FCJ-156 Hacking the Social: Internet Memes, Identity Antagonism, and the Logic of Lulz. *The Fibreculture Journal*, (22 2013: Trolls and The Negative Space of the Internet).

Ryan M Milner. 2016. *The world made meme: Public conversations and participatory media.* MIT Press.

Nir Noon Nave, Limor Shifman, and Keren Tenenboim-Weinblatt. 2018. Talking it personally: Features of successful political posts on facebook. *Social Media+ Society*, 4(3):2056305118784771.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

Limor Shifman. 2019. Internet memes and the twofold articulation of values. *Society and the Internet: How Networks of Information and Communication are Changing Our Lives*, page 43.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XL-Net: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. ”Bidirectional Long Short-Term Memory Networks for Relation Classification”. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China, October.