SarcasmDet at SemEval-2021 Task 7: Detect Humor and Offensive based on Demographic Factors using RoBERTa Pre-trained Model

Dalya Faraj
Department of Computer Science
Jordan University of Science
and Technology
Irbid, Jordan
dfalnore18@cit.just.edu.jo

Malak Abdullah
Department of Computer Science
Jordan University of Science
and Technology
Irbid, Jordan
mabdullah@just.edu.jo

Abstract

This paper presents one of the top winning solution systems for task 7 at SemEval2021, ”HaHackathon: Detecting and Rating Humor and Offense”. The shared task 7 consists of two parts, task-1 with three sub-tasks 1a,1b, and 1c, and task-2. The goal of task-1 is to predict if the text would be considered humorous or not, then if it is yes, predict how humorous it is and whether the humor rating would be perceived as controversial. The goal of task-2 is to predict how the text is considered offensive for users in general. The proposed solution, SarcasmDet, has been developed using RoBERTa pre-trained model with ensemble techniques. The paper describes the submitted system's architecture with the experiments and the hyperparameter fine-tuning that led to this robust system. Our model ranked third and fourth places out of 50 teams in tasks 1c and 1a with F1-Score of 0.6270 and 0.9675, respectively. At the same time, the model ranked one of the top 10 models in task 1b and task 2 with RMSE scores of 0.5446 and 0.4469, respectively.

1 Introduction

In our daily life, the obstacles and difficulties in dealing with sarcasm, bullying, or even abuse of all kinds and ways are increasing day by day (Sheehan et al., 1999; Cleary et al., 2009; Tucker and Maunder, 2015; van Verseveld et al., 2021). Technically, sarcasm and bullying are among the most complex and challenging topics that major companies and institutes seek to address. Artificial intelligence and text processing techniques are the most potent current methods for detecting these problems within texts and images. Sarcasm and abuse are associated with attacking a specific person or group of people either through an unintended joke or, in many cases, by directly affecting the target's psyche. Irony and offensiveness are characterized by their vocabularies that are peppered with humor to conceal the opposite (Lee and Katz, 1998).

Task 7 at SemEval-2021, ”HaHackathon: Detecting and Rating Humor and Offense”, provides two main tasks: task-1 with three sub-tasks (1a,1b, 1c) and task-2. The goal of task-1 is to predict if the text would be considered humorous or not, and if it is yes, then expect how funny it is and whether the humor rating would be perceived as controversial. The goal of task 2 is to predict how the text is considered offensive for users in general. Our solution, SarcasmDet, has been ranked among the top four teams in two sub-tasks. The proposed approach uses the provided dataset, which contains 10K of row text data. We have experimented with several pre-trained language models using the simple transformers library. It is worth mentioning that using the hard-voting ensemble technique has increased our score remarkably.

The paper is constructed as follows: Section 2 provides the related works. Section 3 and 4 describe the shared task and the provided dataset, respectively. Section 5 describes our system solution. Section 6 shows our experiments. Section 7 provides the results, and finally, the conclusion is in Section 8.

2 Related Works

In recent years, social media's development and growth have motivated the NLP research community to detect Humor and Offensiveness. In 2018, SemEval provided different shared tasks to detect emotions and irony in tweets (Mohammad et al., 2018; Van Hee et al., 2018). The top teams' proposed models mostly used LSTM and word embeddings (Abdullah and Shaikh, 2018; Badaro et al., 2018; Wu et al., 2018). In 2019, SemEval also introduced a shared task to discover offensive language in social media. Researchers in (Liu et al., 2019a)

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 527–533
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

used the dataset of the Offensive Language Identification Dataset (OLID) provided by (Zampieri et al., 2019). They ranked first in the task with an F1 (Macro) score of 0.8286 by applying linear model, LSTM, and BERT pre-trained model. In 2020, one of the shared tasks presented in SemEval was about how to change a chunk of text to make the text funnier. The authors(Mahurkar and Patil, 2020; Shatnawi et al., 2020) applied a pre-trained BERT model with different preprocessing for the presented dataset. This paper presents our solution to task 7 in SemEval2021, to detect humor and offensive simultaneously and explains it in detail.

3 Tasks Description

All subtasks of the SemEval2021 task 7 have different requirements. In this section, we have detailed the description for each task.

3.1 Task 1a Humor Detection

Task1a is a binary classification problem. The text should be classified as humor or not based on the answers of 20 participants to whether the particular text was intended to be funny or not. It is considered funny based on the majority of the participants' responses. Table 1 shows an example of the training dataset for task 1a.

#	Example	is humor
348	A babyś laughter can be the most beautiful sound you will ever hear. Unless itś 3 am. And youŕe home alone. And you dont́ have a baby.	1
6	Trabajo,' the Spanish word for work, comes from the Latin term 'trepaliare,' meaning torture.	0

Table 1: Example for task 1a from the train dataset

3.2 Task 1b Average Humor Score

Task 1b is a regression task; humor rating depends on the classified task 1a arguments. If the text was classified as funny (humor), then a question was raised about the level of humorous in the text on a scale of 1-5. Then, they took the average rating as a label. If not humorous text, they used 0 as a label. Table 2 shows an example of the training dataset for task 1b.

#	Example	humor rating
348	A babyś laughter can be the most beautiful sound you will ever hear. Unless itś 3 am. And youŕe home alone. And you dont́ have a baby.	3.1
15	Balsamic vinegar helps slowing the appearance of ageing signs healthy healthy food health.	0

Table 2: Example for task 1b from the train dataset

3.3 Task 1c Humor Controversy

Task 1c is a binary classification problem task; humor controversy depends on the classified arguments from task 1a. If the text was classified as funny (humor), then the task should determine whether the classification of the humor is controversial (1) or not (0). Table 3 shows an example of the training dataset for task 1c.

#	Example	humor controversy
348	A babyś laughter can be the most beautiful sound you will ever hear. Unless itś 3 am. And youŕe home alone. And you dont́ have a baby.	1
8000	Each ounce of sunflower seeds gives you 37% of your daily need for vitamin E vitamin health.	0

Table 3: Example for task 1c from the train dataset

3.4 Task 2 Average Offensiveness Score

Task2 is a regression problem task. The question was asked to determine whether the text is offensive in general, and how much general offensive is between 1-5. Table 4 shows an example of the training dataset for task 2.

4 Dataset Description

The dataset provided by (Meaney et al., 2021) SemEval 2021 organizers for task7 contains 10,000 rows of text data and four columns of labels. The

#	Example	offense rating
27	How do the Chinese select their baby names? They chuck a tin can down the stairs Ping Wong ching Pang	3.8
1498	Today, I overslept and completely missed my 2nd nap.	0

Table 4: Example for task 2 from the train dataset

dataset is divided into three phases: training, development, and evaluation phase datasets. The dataset was collected by surveying US English native speakers of various ages between 18-70 and different genders, political situations, and income levels. The training set contains 8,000 rows of texts with four labels, every text in the data set have been classified based on four questions that were asked to the participants, and each question is related to a specific task. Each of the development set and the test set contain 1000 texts.

4.1 Data Preprocessing

There was no need to implement preprocessing methods for the dataset of task1a and task2. However, the dataset for task1b and task1c contain null values. Therefore, we attempted to convert all null values into zeros, which lowered the data's quality. Therefore, we used another technique, which is dropping the records with null. The later technique increased the data quality and gave better performances.

5 Systems Description

In our solution, we have used the pre-trained language model, RoBERTa (Liu et al., 2019b), that uses a robustly optimized NLP method to improve the Bidirectional Encoder Representations from Transformers. We have also used the BERT pre-trained model (Devlin et al., 2018). RoBERTa is built based on BERT's language masking strategy, which learns to predict knowingly hidden sections of text within unannotated language examples. We have chosen RoBERTa pre-trained model because of the significant improvements in the performance by tuning the BERT training procedure and the architecture based on BERT-large. We have experimented with several deep learning models. In our best-performed solution system, we implemented ensemble technique (Chou et al., 2009) on the best-scored models that include RoBERTa-large, BERT-large trained on 24-layer, 1024 hidden, 16-heads, 355M parameters. We used RoBERTa model from HuggingFace(Wolf et al., 2019) and simpletransformers pre-trained models. More details about each subtask are as follows.

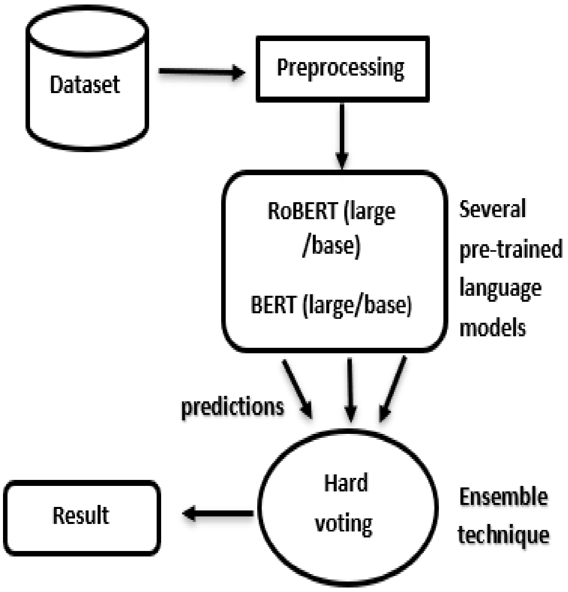

Figure 1: The architecture of our model.

5.1 Task 1a

We have applied RoBERTa (base/large) with different hyperparameters. Then, we have utilized the hard-voting ensemble technique to produce the best model that predicts the label in the test dataset. Our approach's best result has scored 0.9513 F-score in the development phase and 0.9675 F-score in the test phase. The learning rate=1e-5, manual seed= 17, train batch size= 16, and num train epochs= 5.

5.2 Task 1b

In this sub-task, we have applied BERT(base/large) cased with different hyperparameter. Then applied the hard-voting ensemble technique for the best model to predict the label in the test set (learning rate=1e-5, manual seed= 17, train batch size= 16, and num train epochs= 5).

5.3 Task 1c

In this sub-task, we have applied several pre-trained NLP models, such as BERT(base-large), XlNet(large), and RoBERTa(large), but the best solution was obtained from the two previous sub-tasks (task 1a, task 1b). We used the best results of task 1a and task 1b to predict task 1c. If the result from task1a is 1, then that indicates it is humor. If the value of the Humer_Ratting is equal or more than 3, then we consider that humor_controversy to be 1. Otherwise, we assume humor_controversy is 0.

5.4 Task2

We have applied RoBERTa (large/base) with different hyperparameters. Then, we have used the hard-voting ensemble technique for the best model to predict the label in the test dataset (learning rate=1e-5, manual seed= 17, train batch size= 16, and num train epochs= 5).

6 Experiments

We have experimented with several pre-trained NLP models to detect Humor and Offensive through the development and evaluation phases. The pre-trained models include BERT(base/large) that is developed by Google researchers. Also, AlBERT(base/large)(Lan et al., 2019), which is a lite version of BERT to reduce parameters and increase the model speed by reducing memory consumption. Another pre-trained model is XlNet(base/large)(Yang et al., 2019), which introduced the automatic regressive pre-training method and outperformed BERT model in several tasks sentiment analysis, question answering and others. Finally, RoBERTa model, which outperformed most of the pre-trained models, if not all. We have implemented our experiments on google Colab using CPU and GPU. Using collab GPU increased the speed of the experiments by 100%. We used simple transformers library with various hyperparameters, learning rate=1e-5, manual_seed= 17, train_batch size=8-16-32 and epochs= 2-3-5. Our best results accomplished on all tasks was using hard-voting ensemble technique on top of best-scored results by RoBERTa-large and BERT-large-cased. The use of hard-voting technique increased the performance, and the accuracy, remarkably. In the development phase, our model ranked first place in three task1a, task1c, task2, and second place in task1c. However, for the evaluation phase, we have ranked 4th place in task1a, and 3rd place in task1c 3rd. Table 5 shows details of all hyperparameters used on all models for two phases.

Model	**Epoch**	**Batch Size**	**LR**	**Manual Seed**
RoBERTa-large	3,5	8,16,32	1e-5	17
RoBERTa-base	3,5	8,16	1e-5	17
BERT-base	2,3,5	8	1e-5	11,17
BERT-large	3,5	8,16,32	1e-5	11,17
XLNet-base	2,3	8,16	1e-5	17
XLNet-large	3,5	16	1e-5	17
AlBERT-base	2,3	8,16	1e-5	17
AlBERT-large	3,5	16	1e-5	17

Table 5: Hyperparameter was used in all experiments for two phases(development and evaluation) of all tasks.

7 Results

Our solution system results are divided into two phases development and evaluation phase. We experienced several pre-trained language models (RoBERTa, BERT, ALBERT, and XLNET) and implemented them using a simple transformers library in the development phase. In the evaluation phase(test phase), we improved our system solution's capabilities by using different hyperparameters with ensemble techniques. In the following sub-sections, we provided in detail all the results of the evaluation phase.

7.0.1 Task 1a

In task1a RoBERTa-large model outperformed all models with 0.9669 F-score and 0.9590 accuracies. We used the hard-voting ensemble technique to improve our results using the top best five achieved scores by RoBERTa-large and RoBERTa-base model with different hyperparameters. We have increased our solution performance and accomplished 4th place with a 0.9675 f-score and 0.9600 accuracies using this method. Table 6 shows the ensemble results and the top best results for RoBERTa model.

7.0.2 Task 1b

In task1b BERT-large-cased outperformed all models with 0.5468 RMSE. We improved the result

#	model	LR	Batch Size	F-Score	Accuracy
(1)	RoBERTa-base	1e-5	16	0.9654	0.9570
(2)	RoBERTa-base	1e-5	8	0.9660	0.9580
(3)	RoBERTa-large	1e-5	32	0.9661	0.9580
(4)	RoBERTa-large	1e-5	16	0.9663	0.9580
(5)	RoBERTa-large	1e-5	8	0.9669	0.9590
(6)	*	*	*	**0.9675**	**0.9600**

Table 6: Top best experiments used for task 1a in the evalution phase by RoBERTta(large/base) model.*Ensemble for 1,2,3,4,5

with the same method of ensemble and hyperparameters used in the previous sub-task. We have used the top best four achieved scores by BERT-large-cased and BERT-base-cased model, and we achieved 10th place with 0.5446 RMSE. Table 7 shows ensemble results and the top best results for the BERT model.

#	model	LR	Batch Size	RMSE
(1)	BERT-base	1e-5	8	0.5492
(2)	BERT-base	1e-5	16	0.5475
(3)	BERT-large	1e-5	16	0.5468
(4)	BERT-large	1e-5	8	0.5498
(5)	*	*	*	**0.5446**

Table 7: Top best experiments used for task 2 in the evaluation phase by BERT(large/base). * Ensemble for 1,2,3,4

7.0.3 Task 1c

In task1c, we have implemented a method consisting of top of the best task1a and task1b results and using it. We accomplished third place with a 0.6270 F-score and 0.4699 Accuracy.

7.0.4 Task 2

Task2 RoBERTa-large outperformed all other models with 0.4559 RMSE. We have used hard-voting ensemble technique and various hyperparameters with the top five best results by RoBERTa-large and RoBERTa-base. Using this technique, we acheived 10th place with 0.4469 RMSE. Table 8 shows ensemble results and top best results.

#	model	LR	Batch Size	RMSE
(1)	RoBERTa-base	1e-5	8	0.4828
(2)	RoBERTa-large	1e-5	32	0.4741
(3)	RoBERTa-large	1e-5	16	0.4609
(4)	RoBERTa-large	1e-5	8	0.4559
(5)	*	*	*	**0.4469**

Table 8: Top best experiments used for task 2 in the evaluation phase by RoBERTta(large/base) model. * Ensemble for 1,2,3,4

7.1 Error Analysis

Our model was able to predict well in task 1a with an F1-score of 0.9675, but in task 1c, the prediction decreased with an F1-score of 0.52. Figures 2 and 3 show the confusion matrix for tasks 1a and 1c. The reason for this is due to the distribution of the datasets. In task 1a, the dataset was balanced, but in task 1c, the dataset was imbalanced as it contained null values.

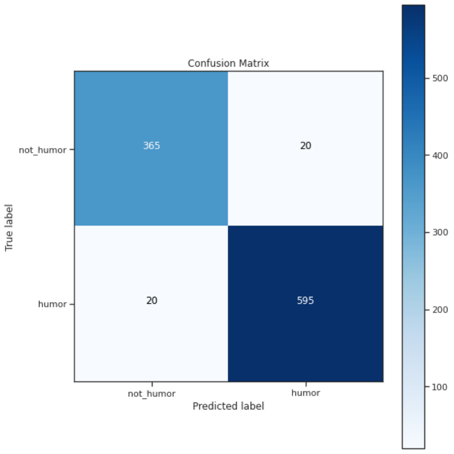

Figure 2: confusion matrix for task 1a.

8 Conclusion

This paper presents and describes our solution system for the SemEval2021 Task7: HaHackathon

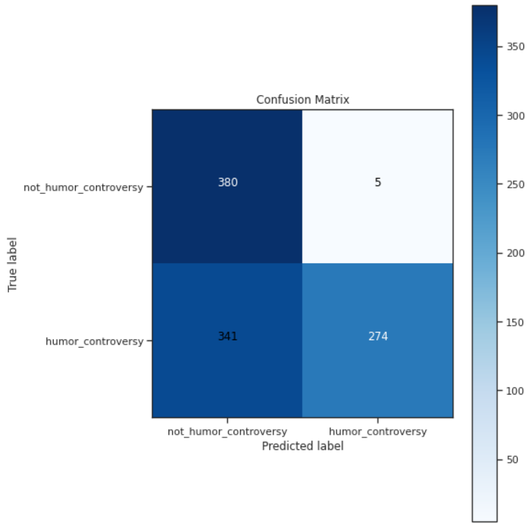

Figure 3: confusion matrix for task 1c.

detecting and rating humor and offense. We have applied several pre-trained language models, such as RoBERTa, BERT, ALBERT, and XLNET, with hard-voting ensemble technique to detect humor and offense mechanism. Our final solution was based on BERT-large-cased model and RoBERTa-large model, which showed remarkable improvements and a high overall outperformance. Our solution system ranked 4th place in task1a with a 0.9675 F-score, 10th place in task1b with a 0.5446 RMSE, 3rd place in task1c with a 0.6270 F-score, and 10th place in task2 with a 0.4469 RMSE.

References

Malak Abdullah and Samira Shaikh. 2018. Teamuncc at semeval-2018 task 1: Emotion detection in english and arabic tweets using deep learning. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 350–357.

Gilbert Badaro, Obeida El Jundi, Alaa Khaddaj, Alaa Maarouf, Raslan Kain, Hazem Hajj, and Wassim El-Hajj. 2018. Ema at semeval-2018 task 1: Emotion mining for arabic. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 236–244.

Te-Shun Chou, Jeffrey Fan, Sharon Fan, and Kia Makki. 2009. Ensemble of machine learning algorithms for intrusion detection. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 3976–3980. IEEE.

Michelle Cleary, Glenn E Hunt, Garry Walter, Michael Robertson, et al. 2009. Dealing with bullying in the workplace: Toward zero tolerance. *Journal of Psychosocial Nursing and Mental Health Services*, 47(12):34–41.

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*.

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

Christopher J Lee and Albert N Katz. 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and symbol*, 13(1):1–15.

Ping Liu, Wen Li, and Liang Zou. 2019a. Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Siddhant Mahurkar and Rajaswa Patil. 2020. Lrg at semeval-2020 task 7: Assessing the ability of bert and derivative models to perform short-edits based humor grading. *arXiv preprint arXiv:2006.00607*.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7, hahackathon, detecting and rating humor and offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Fara Shatnawi, Malak Abdullah, and Mahmoud Hammad. 2020. Mlengineer at semeval-2020 task 7: Bert-flair based humor detection model (bfhumor). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1041–1048.

Michael Sheehan, Michelle Barker, and Charlotte Rayner. 1999. Applying strategies for dealing with workplace bullying. *International journal of manpower*.

Emma Tucker and Rachel Maunder. 2015. Helping children to get along: teachers' strategies for dealing with bullying in primary schools. *Educational Studies*, 41(4):466–470.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Marloes DA van Verseveld, Minne Fekkes, Ruben G Fukkink, and Ron J Oostdam. 2021. Teachers' experiences with difficult bullying situations in the school: An explorative study. *The Journal of Early Adolescence*, 41(1):43–69.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.