CLASSIFICATION OF MENTAL ILLNESSES ON SOCIAL MEDIA USING ROBERTA

Ankit Murarka*
IBM / Raleigh, NC
ankit.murarka1@ibm.com

Balaji Radhakrishnan*
balag59@gmail.com

Sushma Ravichandran*
IBM Research / Yorktown heights, NY
sushma.ravichandran@ibm.com

ABSTRACT

Given the current social distancing regulations across the world, millions suffering from mental illnesses feel isolated and are unable to receive assistance in person. They have increasingly turned to online platforms to express themselves and to look for guidance in dealing with their illnesses. Keeping this in mind, we propose a solution to classify mental illness posts on social media thereby enabling users to seek appropriate help. In this work, we classify five prominent kinds of mental illnesses- depression, anxiety, bipolar disorder, ADHD and PTSD by analyzing unstructured user data on Reddit. In addition, we share a new high-quality dataset[1] to drive research on this topic. Our model is trained on Reddit data but is easily extensible to other social media platforms as well as demonstrated in our results. We also demonstrate how we stress-test our model using behavioral testing. We hope that this work contributes to the public health system by automating some of the detection process and alerting relevant authorities about users that need immediate help.

1 INTRODUCTION

During these unprecedented times when the world is plagued by COVID19, a large number of people have been showing symptoms of clinical anxiety or depression[2]. This can be attributed to a myriad of reasons including lock down, mandatory social distancing, higher unemployment, economic depression and work-related stress. In a report published earlier this year, the American Foundation for Suicide Prevention found that people experience anxiety (53%) and sadness (51%) more often now than before the coronavirus pandemic.

People actively partake in sharing their day to day activities, experiences, feelings, opinions, hopes, desires, and emotions online. These texts provide information which can be used to identify the mental health individuals. Furthermore, the current state of enforced social distancing and isolation has propelled more people to express their emotions on social media as it provides them with an accessible platform to share their thoughts with others, many a times, in search for help.

Due to the paucity of adequate annotated and structured user data in this domain, we decided to generate our own dataset by crawling subreddits on reddit.com[3].

In this work, we use a RoBERTa (Liu et al., 2019) based classifier. RoBERTa's effective performance on unstructured data and its ability to learn contextual information compelled us to explore its power to categorize online user generated texts into various classes of Mental Illness.

* These authors contributed equally

[1] https://github.com/amurark/mental-health-detection

[2] https://afsp.org

[3] https://www.reddit.com

We identify five broad classes of mental illnesses - depression, anxiety, bipolar disorder, ADHD (Attention Deficit Hyperactivity Disorder), PTSD (Post Traumatic Stress Disorder) and an additional 'None' class (which does not pertain to any mental illness). Based on our experiments, we present encouraging results that demonstrate that social media data has the potential to complement standard clinical procedures in the prognosis of mental health.

2 RELATED WORK

Most of the recent such research has revolved around Reddit data: Kim et al. (2020), Gkotsis et al. (2017), Sekulic & Strube (2019), Zirikly et al. (2019). Prior to this recent shift to Reddit data, a lot of the earlier research was focused on utilizing Twitter data: Orabi et al. (2018), Benton et al. (2017), Coppersmith et al. (2015).

There have been a wide variety of approaches ranging from classical NLP techniques to neural network based deep learning methods. Coppersmith et al. (2015) used character level language models to examine how likely a sequence of characters is to be generated by a user with mental health issues. Benton et al. (2017) evaluated a standard regression model, a multilayer perceptron single-task learning (STL) model, and a neural MTL model on detecting multiple types of mental health issues. Orabi et al. (2018) utilized word embeddings in tandem with a variety of neural network models like CNNs and RNNs to detect depression. Gkotsis et al. (2017) experimented with Feed Forward Neural Networks, CNNs, SVMs and Linear classifiers to perform binary classification on mental health posts. Sekulic & Strube (2019) came up with the approach of using Hierarchical Attention Networks(HANs) to detect a wide range of mental health issues like Depression, ADHD, Anxiety etc. and trained a binary classifier for each of the disorders. The most recent work on this was by Kim et al. (2020) who proposed a CNN-based classification model. Once again though, each disorder had its own separate binary classifier to perform the detection.

3 DATASET

The Reddit API was used to crawl 13 Reddit Subreddits for a total of 17159 posts (text and title) to obtain the data for this work. Even though there are a lot of mental illnesses that need addressing, only 5 of them had sufficient data for our purposes. They are: `bipolar`, `adhd`, `anxiety`, `depression` and `ptsd`. The posts in these subreddits were assigned a class label corresponding to the name of the mental illness they were associated with. All the remaining subreddits were carefully chosen as to minimize any chances of thematic content overlap between them and the illness classes. The text from these subreddits were combined and assigned the class label None. While we are currently treating this task as a multi-class classification problem, we duly acknowledge the fact that mental illnesses are generally co-related and require multi-label classification techniques.

While collecting data, we ensured that the number of upvotes for each post in all subreddits is more than 10. We also set a minimum post token length of 30 tokens. These numbers were chosen after carefully perusing through the data in order to retain quality in the dataset. The dataset was preprocessed to remove any URLs or usernames that could potentially contain sensitive information. This was done keeping in mind that the dataset will be released publicly for the purpose of extending this research work.

We also certified that the general topic subreddits did not have a high similarity with the posts corresponding to other the other 5 subreddits. This was done to ensure that we do not have any false negatives while assigning truth labels. In addition to this, we compared the cosine similarity between some of the highest/lowest posts of mental illness subreddits and the general topic subreddits and manually compared the results to find that this distance was higher than the distance between two posts of the mental illness subreddits.

4 MODEL

We propose a RoBERTa (Liu et al., 2019) based multi-class classifier. In addition, we also compare the proposed model against LSTM (Hochreiter & Schmidhuber, 1997) and BERT (Devlin et al., 2018) baselines to demonstrate the superiority of our approach.

Models	**posts**				**titles**				**posts+titles**			
	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
LSTM	0.74	0.72	0.72	0.72	0.65	0.64	0.64	0.64	0.77	0.76	0.76	0.76
BERT	0.83	0.82	0.82	0.82	0.72	0.71	0.71	0.71	0.87	0.87	0.87	0.87
RoBERTa	0.86	0.86	**0.86**	0.86	0.73	0.72	**0.72**	0.72	0.89	0.89	**0.89**	0.89

Table 1: Results: Classification Report

Class	**posts**			**titles**			**posts+titles**		
	P	R	F1	P	R	F1	P	R	F1
adhd	0.87	0.88	0.87	0.77	0.79	0.78	0.91	0.92	0.91
anxiety	0.78	0.83	0.81	0.69	0.64	0.67	0.87	0.85	0.86
bipolar	0.88	0.79	0.83	0.58	0.63	0.60	0.88	0.83	0.86
depression	0.77	0.83	0.80	0.65	0.78	0.71	0.81	0.88	0.84
ptsd	0.88	0.85	0.86	0.75	0.62	0.68	0.88	0.89	0.88
none	0.99	0.95	0.97	0.94	0.88	0.91	1.00	0.98	0.99

Table 2: Results: RoBERTa Class-wise results

We chose a sequence length of 35 for titles, and 512 for posts and posts+titles. The models were fine-tuned for 10 epochs with a learning rate of 1e-5 and Adam (Kingma & Ba, 2014). A batch size of 32 was used while fine-tuning on the titles whereas a batch size of 16 was utilized to fine-tune on posts and posts+titles. Cross-entropy was used as the loss function. A dropout (Srivastava et al., 2014) layer with probability of 0.3 was used for regularization.

5 Result Analysis

The results from our experiments are documented in Tables 1 through 3. As can be observed from Table 1, our proposed RoBERTa based classifier far outperforms the baseline LSTM in all categories. The BERT classifier has results which are quite close to that of RoBERTa's and both beat LSTM by a significant margin, showcasing the incredible capabilities of pre-trained Transformer based architectures. In fact, our RoBERTa model fine-tuned on just the titles was able match the performance of the LSTM model trained on posts. The RoBERTa model was able to achieve an F1 score of 0.86 on the posts and 0.89 on posts+titles which are extremely promising given the complex nature of the multi-class mental illness classification task. The jump in accuracy between posts and posts+titles is not as drastic as the jump between titles and posts. This indicates that the posts offer far more valuable information when compared to the titles and also the fact that most of the useful and relevant information can be extracted from the posts alone. This strong performance on just the posts bodes well for the extensibility of our approach as this can be applied on almost any given social media post without the need for structure in the data like titles, user names, user history, etc.

The rest of this section will focus solely on the results of our best performing RoBERTa model. Table 2 showcases the granular class-wise results of the RoBERTa model. The first strikingly obvious result is the high accuracy with which the model is able to detect non-illness related posts. Even with just the titles, the model is able to classify the `none` class with an f1 score of more than 0.9. When using posts, just 3 illness related posts across the entire test dataset were misclassified as non-illness posts. This number further reduces to 0 when using titles+posts. This shows that the model will detect mental illness posts correctly nearly every single time.

When it comes to the class wise performance amongst the mental illnesses, the two best performing classes are `adhd` and `ptsd` whereas the two worst performing classes are `depression` and `anxiety`.

The performance of `depression` and `anxiety` classes can be attributed to a few factors. The average number of words per post for `depression` and `anxiety` are the least for any given class. For instance, `depression` posts have roughly 53% lesser textual data when compared to `ptsd`

[4]In all our tables, P and R stand for Precision and Recall respectively

Input	Actual	Predicted
often times i'll get distracted from my thoughts either by external influences or just another idea coming in, and then i have to spend a good 5 minutes trying to work out what i was thinking about again.	adhd	adhd
once i come down from flashbacks or panic attacks, i get really bad disassociation. sometimes lasting for days. does anyone else go through this. any tips on how to stop it. i tried grounding but im so far gone it doesn't help.	ptsd	ptsd
i can't sit still when i get my eyebrows done, and when i'm in class i usually doodle to focus. i pay attention very well in school regardless of that, and drawing helps me focus.	anxiety	adhd
i'm flying from dallas to hong kong in january and it's 17 hours. i've flown 12-13 hour flights before and they really mess with me. so i'm wondering - what are your tips for not going crazy on such a long flight? ps: i'm terrible at sleeping on planes. thinking about taking some sleepy meds to see if it'll help	none	anxiety

Table 3: Results: Interesting Examples

posts. In addition, studies show that depression might often occur in tandem with another mental illness and our data and results back this up as well. The `depression` word occurs in 12% of `anxiety` posts, 12% of `ptsd` posts and 31% of `bipolar` posts. Similarly, `anxiety` occurs in 20% of `ptsd` posts, 12% of `adhd` posts and 14% of `bipolar` posts. This implies that the model cannot give high importance to the mention of these class names like it can with rest of the illnesses, thus making the classification of these 2 classes that much harder.

This can also explain the relatively lower precision scores(higher number of False Positives) for `depression` and `anxiety`. When the other illnesses(excluding `depression` and `anxiety`) are misclassified, they are almost always misclassified as either `depression` or `anxiety`. In the same figure, we can see that `depression` and `anxiety` are often misclassified as each other due to the reason that they commonly occur together.

There are more posts in the `adhd` and `ptsd` classes that mention the words `depression` and `anxiety` than their respective class names itself. One would assume that this would result in sub-par results, but, these classes actually perform the best. This really showcases the true potential of our model, where it doesn't just rely on mention of class names, but has a strong understanding of the context of the post itself. Additionally, the symptoms or descriptions provided for these classes could be strong, unique and discriminative enough for the model to be able to classify them correctly even with all the mentions of other class names.

In Table 3 we have documented a few interesting results we observed in the test set. In the first two examples on the table, the RoBERTa model was able to classify the posts correctly without the presence of class names in the input. The prediction is based purely on contextual information learnt about the class labels during the training process. The next two results are interesting because the truth label assigned to the input text may or may not correspond to actual mental illness described in the text. Since, we are not domain experts ourselves, we would need expert intervention to substantiate this theory. As a part of future work, getting professionals to annotate our dataset might help strengthen the model for such examples. Further samples from our qualitative testing on various social media platforms can be found on our accompanying web-page[5].

6 Conclusion and Future Work

Our chief motivation behind this work is the current pandemic and mandatory confinement world-wide. We believe that social media has become the prime mode of communication for people and has paved way for them to vent freely without judgement.

Our roadmap includes getting all our data annotated by mental health experts in order to verify our annotations. This would also assist us in creating a multi label dataset which is more representative of this problem when compared to a multi class one. Our work involves two kinds of texts- long

[5]https://mental-health-classification.github.io/

and short - both of which are common to the internet community. Hence, our work can easily be extended to many websites.

In conclusion, we believe that our work explores an interesting line of research where NLP is used to bridge the gap between virtual and real life of users and help those in need of medical attention.

REFERENCES

Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text. 2017. URL https://www.aclweb.org/anthology/E17-1015v2.pdf.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, 2015. URL https://www.aclweb.org/anthology/W15-1204.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7(1):1–10, 2017.

Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. 1997.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10(1):1–6, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019. URL https://arxiv.org/abs/1907.11692.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, 2018. URL https://www.aclweb.org/anthology/W18-0609.pdf.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. URL https://arxiv.org/abs/2005.04118.

Ivan Sekulic and Michael Strube. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 322–327, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5542. URL https://www.aclweb.org/anthology/D19-5542.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, January 2014. ISSN 1532-4435.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, 2019. URL https://www.aclweb.org/anthology/W19-3003.pdf.

Synonym Replacement					**Label Removal**				
Test Set Modified	**posts**				**Test Set Modified**	**posts**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.86	0.85	0.85	0.85	10%	0.85	0.84	0.84	0.84
50%	0.85	0.84	0.84	0.84	50%	0.81	0.80	0.80	0.80
100%	0.83	0.83	0.83	0.83	100%	0.75	0.74	0.75	0.74
Test Set Modified	**titles**				**Test Set Modified**	**titles**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.73	0.72	0.72	0.72	10%	0.72	0.71	0.71	0.71
50%	0.71	0.71	0.71	0.71	50%	0.67	0.67	0.67	0.67
100%	0.68	0.67	0.67	0.67	100%	0.61	0.61	0.60	0.61
Label Replace: 'illness'					**Label Replace: random**				
Test Set Modified	**posts**				**Test Set Modified**	**posts**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.84	0.83	0.84	0.83	10%	0.83	0.82	0.83	0.8
50%	0.78	0.77	0.77	0.77	50%	0.71	0.71	0.71	0.71
100%	0.70	0.67	0.68	0.67	100%	0.58	0.57	0.57	0.57
Test Set Modified	**titles**				**Test Set Modified**	**titles**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.72	0.71	0.71	0.71	10%	0.71	0.71	0.71	0.71
50%	0.67	0.65	0.65	0.65	50%	0.64	0.64	0.64	0.64
100%	0.62	0.57	0.58	0.57	100%	0.54	0.54	0.54	0.54

Table 4: Behavioral Tests

A Appendix

A.1 Behavioral Testing

Although classification metrics are generally regarded sufficient in estimating the performance of Bert-based models, a recent inclination of NLP researchers to perform behavioral testing inspired us to stress test our models as well.

For all our tests, we used our proposed RoBERTa model and applied these tests to inputs that were either titles or posts. Since we hope to extend our model to other social media platforms, we do not always expect input texts to have a title as well as a descriptive text/post. We adopted the Checklist approach (Ribeiro et al., 2020) which involve tests conducted to comprehensively analyze the model's performance.

A.1.1 Synonym Replacement

Synonym replacement is a kind of Invariance Test where label-preserving perturbations are made to the test set. As labels, the root form of the mental illnesses was chosen- `depress`, `ptsd`, `anxiou/anxiet`, `bipolar` and `adhd`. Python's NLTK package and WordNet were used for these tests.

This test is conducted such that the root words are not perturbed when modifying the test set. For each post, 10% of the tokens were randomly selected (not including the stop words or the root words). Each token was then replaced with one of its synonyms. We used the same logic for titles. We set a max and min on the number of tokens to be selected for replacement - this was (4, 30) for posts and (1, 5) for titles. Since each token was replaced with a synonym, the class label for the samples was not changed. We did this for 10, 50 and 100 percent of the test set and observed results.

In all three cases, we expect the classification metrics to drop. For the case when 10% of the test case was modified the drop was much lower as compared to when 100% of the test case was modified. The results are documented in Table 4. When comparing these results to those in Table 2 we find that the drop in each category is about 2-4% for posts and 5-7% for titles. The lower drop can be attributed to the fact that synonym replacement does not alter the semantics of the input text. Therefore the model was able to draw sufficient information from the input.

Subreddit	Number of posts	Average no. of words (posts)	Average no. of words (titles)	Average Upvotes	Highest Upvotes	Lowest Upvotes
r/depression	3062	152.74	12.20	517.19	4802	11
r/anxiety	3027	170.38	11.75	246.07	3349	11
r/ptsd	2501	233.55	10.14	38.4	443	11
r/adhd	3082	198.55	13.71	377.13	4484	11
r/bipolar	3009	203.28	9.26	32.37	363	11
none	2478	238.52	15.76	6715.33	199295	11

Table 5: Dataset: Statistics

A.1.2 MASKING

We also performed a Directional Expectation test on the model. This is similar to the previous test but is instead performed only on labels. The labels, as defined in the previous subsection, are a list of the root form of mental illness class labels. We noticed that the root words appear often in our input texts. This behavioral test was performed to observe our model's dependency on these words. For all the tests below, we modified only those tokens that contained a root word.

In the first case, for every post from a subreddit related to a mental illness, the root form of its class label was removed from the input. For example, the input text: *I feel happy for some time and then depressed again. I'm definitely bipolar* from the `r/bipolar` subreddit, was modified to *I feel happy for some time and then depressed again. I'm definitely*. Note that changes were not made to the word *depressed* in the input. The class label for each modified sample was not changed after the perturbations. Like the previous subsection, these tests were performed on 10, 50 and 100% of the test set.

In the second case, instead of entirely removing the tokens, we replaced it with a generic token `illness`. We expected this modification to retain some semantic information that was lost in the previous test. However, we found that adding a generic token introduced some noise which reduced the overall performance of the model.

Lastly, the tokens were replaced by a randomly chosen root form of a mental illness other than its class label. With this test, we expect to force the model to pick between the label and non label tokens during classification. We believe that this is an interesting scenario to observe.

In all three cases (Table 4), the model performance drops by some degree when compared to Table 2. The first two cases showed a somewhat similar performance drop. However, the model performance was worse than that of the Synonym Replacement test. This means that the model depends on the existence of the root words in the input text to some degree.

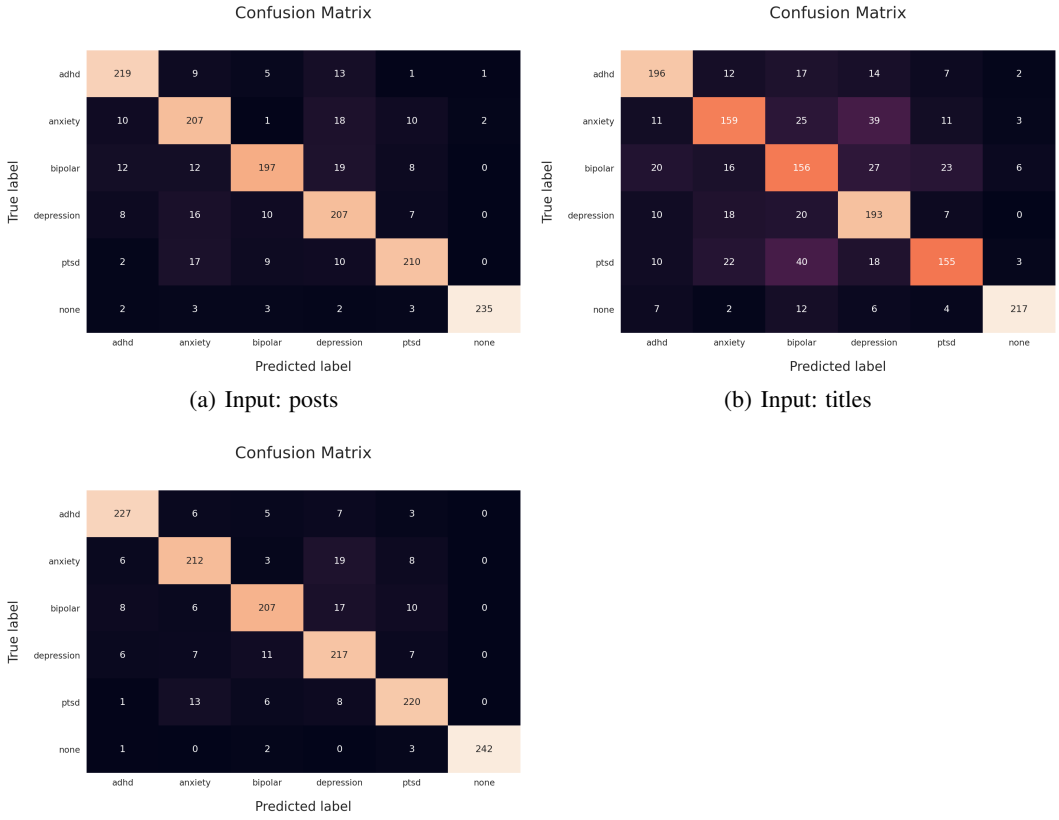
In the third scenario, we note that the performance drop is higher. Although the test is meant to confuse the model, we observed that in some cases (especially for input: posts), we got an F1 of 0.82 with 10% of the modified test and 0.71 with 50% of the modified test. This is only possible if the model gathered sufficient information from the non label text in the input.

A.2 DATASET STATISTICS

Table 5 shows the statistics collected for each subreddit.

A.3 RESULTS: ROBERTA CONFUSION MATRICES

Figure 1 showcases the confusion matrices for the RoBERTa classifier.



(a) Input: posts

(b) Input: titles

(c) Input: posts+titles

Figure 1: RoBERTa: Confusion Matrices