Detecting Moments of Change and Suicidal Risks in Longitudinal User Texts Using Multi-task Learning

Tayyaba Azim[*], **Loitongbam Gyanendro Singh**[*], **Stuart E. Middleton**
School of Electronics and Computer Science,
University of Southampton, Southampton, UK
{ta7g21, gsl1r22, sem03}@soton.ac.uk

Abstract

This work describes the classification system proposed for the Computational Linguistics and Clinical Psychology (CLPsych) Shared Task 2022. We propose the use of multitask learning approach with a bidirectional long-short term memory (Bi-LSTM) model for predicting changes in user's mood (Task A) and their suicidal risk level (Task B). The two classification tasks have been solved independently or in an augmented way previously, where the output of one task is leveraged for learning another task, however this work proposes an 'all-in-one' framework that jointly learns the related mental health tasks. Our experimental results (ranked top for task A) suggest that the proposed multi-task framework outperforms the alternative single-task frameworks submitted to the challenge and evaluated via the timeline based and coverage based performance metrics shared by the organisers. We also assess the potential of using various types of feature embedding schemes that could prove useful in initialising the Bi-LSTM model for better multitask learning in the mental health domain.

1 Introduction

Mental illness has greatly affected a vast majority of world's population due to COVID-19 and its resulting economic recession. According to the world health organisation (WHO), global prevalence of anxiety and depression has increased by a massive 25% raising concerns about providing mental health and psychosocial support to the population as a COVID-19 response plan[1]. Many social media platforms have risen to the challenge by offering space to online users to self report their mental health issues, receive counselling support and resolve their mental health issues. This activity has resulted in two research trends: (1) the surge in development of machine learning algorithms that can automatically detect mental health issues from the language used in social media platforms and (2) the development of new and better diagnostic measures and mental health monitoring tools suitable for the clinical community. Most of the research tasks revolve around classifying individuals on the basis of suicide risk or having a mental health condition (Chancellor and De Choudhury, 2020), however a few have thought of monitoring individual's mood and mental health in real time (Tsakalidis et al., 2022a,b). Despite the growing interest in this interdisciplinary space, there are challenges regarding the availability, use and validity of mental health data gathered from social media platforms and decisions drawn from it.

Table 1: Statistics of the training data set provided for the CLPSych Shared Task 2022.

Moments of Change (Task A)				
Data Set Attributes	None (O)	Escalation (IE)	Switch (IS)	Total
No. of Users	147	87	118	352
No. of Posts	4043	773	327	5143
Avg. No. of Users per post	27.50	8.88	2.77	–
Avg. No. of Words Per Post	75.33	231.82	214.085	–

Suicidal Risk Levels (Task B)				
Data Set Attributes	Low	Moderate	Severe	Total
No. of Users	14	87	103	204
Avg No. of Timelines	1.42	2.17	1.60	–

This paper describes our work identifying moments of change in user's mood (Task A) and suicidal risk level (Task B) in the CLPsych Shared Task 2022. We have experimented with several different sentence and word embedding techniques to draw semantically meaningful features for initialising the multitask sequential model. The model utilised for sequential representation of data is Bidirectional Long Short-Term Memory (Bi-LSTM) (Balikas et al., 2017), trained jointly for multiple tasks (Task A and Task B). The multi task outputs determine

[*]Equal contributions.

[1]https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide (Accessed on 25.5.2022)

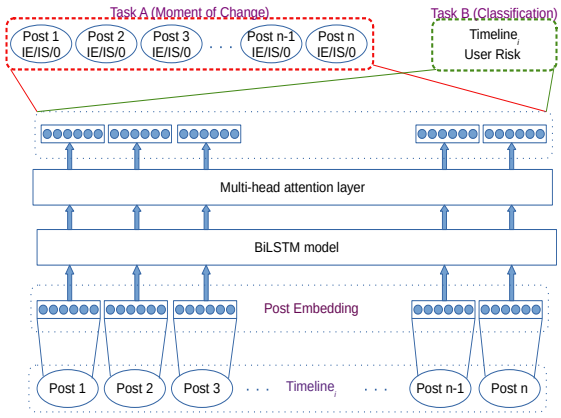

Figure 1: A high-level architecture of the proposed multi-task model for determining moments of change and suicidal risk of users in a particular timeline.

the moments of change in user's mood as well as assess the level of suicidal risk from their posts.

2 Shared Task and Data Set

We have participated in two tasks introduced by the organisers: The first task (*Task A*) is to predict the changes in user's mood over time based on the linguistic content gathered from their posting activity shared on online social media platforms. This is a post-level sequential classification task that aims to detect those sub-periods where a user's mood deviates from their baseline mood. Sequence of an individual's posts over a time span of two months is collected for this shared task (Losada and Crestani, 2016; Losada et al., 2020). The progression in user's mood is categorised as follows: (1) Switch *(IS)*, which signifies a sudden change in user's mood, (2) Escalation *(IE)*, which denotes a gradual shift in user's mood and (3) None *(O)*, denoting no change in user's mood over time. The mood shifting is graded on a scale from positive to negative. This information is further used for *Task B* where user's suicidal risk level is predicted as *Low*, *Moderate* and *Severe* based on the longitudinal mood changes of the user(Shing et al., 2018; Zirikly et al., 2019). The class distribution of the data for each of these labels is shown in Table 1. In order to tackle data imbalance issues, the 'No Risk' and 'Low Risk' label instances were merged and represented as 'Low Risk' examples in the data set for Task B. The task participants were required to sign data use agreements and abide by ethical practice during the competition.

3 Methodology

This section demonstrates the stages involved in developing the proposed multi-task model for determining moments of change in mood and user's suicidal risk determined through a sequence of posts in user's timeline. Figure 1 shows the high-level model architecture for both the tasks.

3.1 Text Preprocessing

The content of the user posts go through several preprocessing steps, including removing stopwords and normalizing keywords (converting to lowercase, removing URL links). Furthermore, the username[2] present in the post is replaced with *@user* to anonymize the mentioned user.

3.2 Semantic Embedding of User Posts

After preprocessing, the user posts are represented using off-the-shelf pre-trained embedding methods to capture the user post's semantics. The pre-trained embedding methods represent the semantics of the posts using fastText word embedding (Bojanowski et al., 2016) and Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). Each post P_i with n tokens can be represented using the pre-trained fastText (FT) word embedding[3] by simply averaging the semantic embeddings of the words present in the post, i.e.,

$$P_i^{FT} = \frac{1}{n}\sum_{pi=1}^{n} \mathbf{w}_{pi}, \mathbf{w}_{pi} \in \mathbb{R}^{300} \quad (1)$$

Recently, the RoBERTa model has yielded much better results in recognizing emotions than other transformer variants such as BERT, XLNet, DistillBERT, and ELECTRA (Cortiz, 2021). Therefore, the RoBERTa-based natural language inference pre-trained model (*'nlirobertalarge'*) is used in addition to fastText embedding to represent the post representation P_i, i.e., $P_i^{SBERT} \in \mathbb{R}^{1024}$.

In order to understand the emotional expressions in text, user's posts are further classified using pre-trained RoBERTa-base model[4] trained on $\widetilde{58}$ million tweets from the TweetEval benchmark (Barbieri et al., 2020) for six different tasks: *emoji*,

[2]Tokens starting with @ symbol.

[3]Pre-trained word embeddings obtained from the Wikipedia corpus https://dl.fbaipublicfiles.com/fastText/vectors-english/wiki-news-300d-1M.vec.zip

[4]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

emotion, *hate*, *irony*, *offensive*, and *sentiment*. The emoji classification task has 20 categories; emotion classification has four; hate, irony, and offensive classification tasks each have two categories; and sentiment classification tasks have three categories. The task-specific scores therefore represent an additional 33 dimensional feature vector to differentiate each user's posts based on the task-specific scores. The post P_i can be represented by aggregating the scores of task-specific pre-trained model ($Scores_t$):

$$P_i^{Score} = \forall_{t \in T} \, Concat(Scores_t(P_i)), P_i^{Score} \in \mathbb{R}^{33} \quad (2)$$

where T is the set of six tasks, i.e., *emoji*, *emotion*, *hate*, *irony*, *offensive*, and *sentiment* and $Concat$ represents the score concatenation for all six tasks.

3.3 Multi-Task Model

A user can post *n* number of posts in a particular timeline t_{ij} ranging from time i to j. The objective of the proposed multi-task model is to predict the moments of change (either IE, IS, O) in the user's posts (Task A) and also classify the suicidal risk of the users (Task B) given the sequence of posts in a particular timeline t_{ij}.

3.3.1 Moments of Change Classification

The problem of predicting the moments of change in the user's mood can be viewed as a sequence tagging problem. The learning model predicts the changes in user's mood for each post sequentially, given the sequence of posts in a timeline. This study proposes to use the bidirectional LSTM (Bi-LSTM) (Zhang et al., 2015) model to capture the sequential information of the user posts in a timeline. The Bi-LSTM model generates dense representation for each post, encoding the sequential information of neighbouring posts in both directions, i.e., the user's previous and subsequent posts. Specifically, the Bi-LSTM model encodes the post sequence representation by concatenating the outputs of two LSTMs, namely LSTM-forward ($LSTM_f$) and LSTM-backward ($LSTM_b$) models. $LSTM_f$ processes the post sequence from left to right, i.e., P_1, P_2, ..., P_n, whereas $LSTM_b$ process the post sequence from right to left, i.e., P_n, P_{n-1}, ..., P_1. Each LSTM model consists of a repeating unit called memory cell, which takes current post, previous hidden state, previous cell state ($\mathbf{x}_t$, $\mathbf{h}_{t-1}$, $\mathbf{c}_{t-1}$) as input and produces current hidden state and cell state information i.e. $(\mathbf{h}_t, \mathbf{c}_t) = LSTM(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1})$. Therefore, the encoded representation of post P_t is generated by concatenating the hidden state information obtained by $LSTM_f$ and $LSTM_b$ outputs, i.e., $\mathbf{h}_t$ = $(\mathbf{h}_t^{(f)} \oplus \mathbf{h}_t^{(b)})$. The whole timeline t_{ij} can be represented as $\mathbf{H_{ij}} \in \mathbb{R}^{n \times d}$ where $\mathbf{H_{ij}}$ is a matrix of the encoded representation of n posts of d dimension[5]. The encoded representation of the posts is then fed to the softmax classifier to predict the user's moment of change, i.e.,

$$\mathbf{Task_a} = Softmax(\mathbf{H_{ij}}\mathbf{W_a}^T + \mathbf{B}) \quad (3)$$

where $\mathbf{W_a} \in \mathbb{R}^{c \times d}$ is the neural weight parameters, c is the three classes of the moment of change categories (i.e., IE, IS, O), and $\mathbf{B} \in \mathbb{R}^{n \times c}$ being the neural network biases.

3.3.2 User Suicidal Risk classification

Using the same encoded representation, the user's risk can be classified for the timeline t_{ij} by flattening the matrix $\mathbf{H_{ij}}$, i.e.,

$$\mathbf{Task_b} = Softmax(flatten(\mathbf{H_{ij}})\mathbf{W_b}^T + \mathbf{B}) \quad (4)$$

where $\mathbf{W_b} \in \mathbb{R}^{r \times nd}$ is the neural weight parameters, r being the number of user risk categories in Task B, and $\mathbf{B} \in \mathbb{R}^r$ being the neural network biases. Further, the user risk can also be classified by embedding an attention layer over the encoded representation $\mathbf{H_{ij}}$ before $flattening$ to give more attention to the user's post that influences the user risk classification decision. The output of the multi-head attention[6] layers generate an attention weighted encoded representation $\mathbf{H_{ij}^a}$ of the same dimension as $\mathbf{H_{ij}}$. The impact of adding an attention layer could be seen in the tables discussed in the results section.

The current model classifies the user's suicidal risk for a particular timeline t_{ij}. However, a user can have multiple timelines $\{t_{ab}, t_{cd}, ..., t_{ij}\}$, hence the user risk must be classified considering all the timelines. Since the model classifies the user risk for each timeline, i.e., $\{\mathbf{Task_b}_{ab}, \mathbf{Task_b}_{cd}, ..., \mathbf{Task_b}_{ij}\}$, the final user risk $\mathbf{Task_b}$ is classified using a simple heuristic approach. The user risk is classified based on the prediction of the user's risk severity level across the timelines, i.e., if the model has predicted *Severe* in one of the timeline then the user is considered to be at *Severe* risk; followed by

[5] 100 LSTM units

[6] 8 heads

Moderate-level and *Low*-level risks. We can also consider a voting method to classify the user risk based on the output of all timelines. This study consider evaluating the user risk classification based on the heuristics of risk severity level.

4 Experiment and Results

In this work we have used two different combinations of feature embeddings for the user posts. For the ease of reference, we consider naming them as P_{emb} which is the concatenation of fastText and SBERT embeddings ($P^{FT} \oplus P^{SBERT}$) and $P_{task-emb}$ which is the concatenation of fastText, SBERT, and task-specific scores of the post ($P^{FT} \oplus P^{SBERT} \oplus P^{Score}$).

Models: The efficacy of the proposed model is evaluated on two types of post embeddings (P_{emb}, $P_{task-emb}$), with and without the attention layers. This, eventually leads us to four different types of models for evaluation: (i) Multitask: model using P_{emb}, (ii) Multitask-score: model using $P_{task-emb}$, (iii) Multitask-attn: model with attention layer using P_{emb}, and (iv) Multitask-attn-score: model with attention layer using $P_{task-emb}$.

Evaluation Metrics: The performance of the proposed model is evaluated using metrics Precision, Recall and F1 Score on the validation set. We also show window-based and coverage-based evaluation metrics (Tsakalidis et al., 2022b) used by the CLPsych organisers to assess the models' performance on the test set.[7]

Implementation Details: The train data set is initially divided into train, validation and test sets using the ratio: 60:20:20, to optimise the Bi-LSTM parameters. Once the parameters are fine tuned using the validation set, we retrain the model again with 80% of the train data and test it on 20% of the unseen test data. After fine tuning, the Bi-LSTM model is trained for 50 epochs with 64 batch size. The maximum sequence length for Bi-LSTM is set to the maximum number of posts in a timeline, i.e., 122 (see Appendix). Categorical cross-entropy loss and Adam optimizer are used to train the model on both the tasks. The implementation was done using Keras API and is available at https://github.com/stuartemiddleton/uos_clpsych.

Table 2 shows the results of our model on the validation set using the standard evaluation metrics.

[7]Please note that the data set is imbalanced and therefore intuitions just drawn from only accuracy are not correct.

Table 2: Performance of the proposed models on Task A and Task B using the validation set.

Model	Moments of Change			Suicidal Risk Levels		
	P	R	F1	P	R	F1
Multitask-attn-score	0.674	**0.800**	**0.724**	**0.415**	**0.397**	0.382
Multitask-score	**0.680**	0.760	0.713	0.355	0.331	0.334
Multitask	0.582	0.717	0.629	0.352	0.327	0.335
Multitask-attn	0.663	0.697	0.676	0.408	0.378	**0.388**

Here, the precision, recall and F1 score values obtained for each class (see Table 5 in the appendix) have been macro-averaged by calculating the arithmetic mean of individual classes' precision, recall and F1 scores. We have used the macro-averaging score to treat all the classes equally for evaluating the overall performance of the classifier regardless of their support values (i.e the actual occurrences of the class in the data set). Here, we observe that **Multitask-attn-score** model gives more promising results as compared to other enlisted models on both tasks. This behaviour is reflected in the classification results on test data too (Table 3), where **Multitask-attn-score** has outperformed the remaining feature embeddings with the Bi-LSTM model as well as the baseline state of the art results (Tsakalidis et al., 2022a). From the model outcomes in Table 2 and 3, one could also see the impact of introducing attention layers in the Bi-LSTM model. Adding attention layers in Bi-LSTM model has helped accuracy for both the tasks.

Given the class imbalance in the data set with majority of post instances belonging to the *None(0)* class and minority instances to *Escalation (IE)* and *Switch (IS)* classes, we see the performance is compromised and biased towards the majority class, i.e. the classifier is more sensitive to detecting the majority class (*None(0)*) patterns precisely but less sensitive to detecting the minority class patterns {*IE*, *IS*}. See Table 5 in the Appendix to observe the precision, recall and F1 score of the models for each individual class in task A. The data distribution is skewed for task B too, thus influencing its results for majority and minority classes shown in Table 6. Overall, on the validation set, the proposed models have shown better recall rate than precision, revealing low false negatives than the false positives.

Table 3 and Table 4 show the performance of our proposed approach with variable feature encoding schemes and attention layers in Bi-LSTM on the test set provided by the CLPsych Shared Task 2022. The entire train set comprising of 5143

Table 3: Performance of the proposed models on Task A using the test set. The traditional post-level, coverage-based, and timeline-based evaluation metrics based on precision (P), recall(R) and F1 score are shown for comparison and analysis with the baseline results (Tsakalidis et al., 2022a).

	Post-level Metrics (Macro average)			Coverage-based Metrics (Macro average)		Timeline-level Metrics (Macro average) Window-1		Window-2		Window-3	
	P	R	F1	P	R	P	R	P	R	P	R
Multitask-attn-score	**0.689**	**0.625**	**0.649**	0.506	**0.503**	**0.676**	**0.652**	0.693	**0.670**	0.708	**0.686**
Multitask-score	0.677	0.595	0.625	0.492	0.467	0.662	0.605	0.681	0.622	0.695	0.632
Multitask	0.680	0.579	0.607	**0.521**	0.441	0.674	0.592	**0.695**	0.608	**0.723**	0.623
Majority	NaN	0.333	0.280	NaN	0.141	NaN	0.333	NaN	0.333	NaN	0.333
TFIDF-LR	0.545	0.495	0.492	0.377	0.424	0.496	0.539	0.505	0.550	0.506	0.551
BERT-TalkLife-Focal	0.522	0.386	0.380	0.260	0.204	0.582	0.392	0.608	0.405	0.608	0.405

Table 4: Performance of the proposed model on Task B using the test set. The precision (P), recall (R), and F1-scores (F1) shown are macro-averaged over the user's risk categories and compared to the baseline results (Tsakalidis et al., 2022a).

	(Macro average)			(Micro average)		
	P	R	F1	P	R	F1
Multitask-attn-score	**0.618**	**0.427**	**0.451**	**0.482**	**0.469**	**0.438**
Majority	0.156	0.333	0.212	0.219	0.468	0.299
TFIDF-LR	0.302	0.338	0.295	0.412	0.468	0.406

posts is used to train the proposed model with the optimal parameters defined above and then its efficacy is assessed on the given test set comprising of 1052 posts. On the test set, the proposed models have shown higher precision than recall. When compared to the baseline results, our submission on task A has topped the ranking results on the test set, whereas for task B we stood second in the shared task based on the timeline based and coverage based metrics.

5 Conclusion and Future Work

This work demonstrates the power of using various feature embeddings for multi task learning with Bi-LSTM on the CLPsych Shared Task 2022 data set. We have tried several different textual embeddings to represent the content of user's posts. These embeddings are passed on to the Bi-LSTM which is trained to learn two labels jointly. The model has shown to give promising results on the test set when attention layer is incorporated and complete set of feature embeddings (fastText+SBERT+TaskScore) is utilised. On Task A, our team topped the post-level classification problem based on the window based and coverage based statistics, whereas for Task B, we showed second best results in the competition.

In future, we would like to compare our proposed model with other single task learning models trained using separate loss functions. Given the correlation between the shared tasks, multi-task learning is expected to yield good results as shown in this paper, however it will be interesting to explore the underlying user information (e.g. age, gender, etc) that could be explicitly added to support tasks for mental health and suicidal risk prediction. Also in order to mitigate the effects of imbalanced classes, we would like to improve our developed pipeline using resampling techniques.

Ethical Statement

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20).

Acknowledgements

The authors are particularly grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, to the annotators of the data for Task A, to the clinical experts from Bar-Ilan University who annotated the data for Task B, the American Association of Suicidology, to NORC for creating and administering the secure infrastructure and providing researcher support and to UKRI for providing funding to the CLPsych 2022 shared task organisers. This work was supported by the Engineering and Physical Sciences Research Council (EP/V00784X/1).

References

Rie Kubota Ando and Tong Zhang. 2005. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *Journal of Machine Learning Research*, 6:1817–1853.

Georgios Balikas, Simon Moura, and Massih-Reza Amini. 2017. Multitask Learning for Fine-grained Twitter Sentiment Analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1005–1008.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv preprint arXiv:2010.12421*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *NPJ Digital Medicine*, 3(1):1–11.

Diogo Cortiz. 2021. Exploring Transformers in Emotion recognition: A Comparison of BERT, Distillbert, Roberta, Xlnet and Electra. *arXiv preprint arXiv:2104.02041*.

David E. Losada and Fabio Crestani. 2016. A test Collection for Research on Depression and Language Use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 272–287. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022a. Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying Moments of Change from Longitudinal User Text. *arXiv preprint arXiv:2205.05593*.

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.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. 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*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Appendices

Tables 5 and 6 show the evaluation metrics by class in Task A and Task B. Figure 2 shows the post distribution in the training set.

Table 5: Performance of the proposed models on Task A using the validation set. The traditional post-level evaluation metrics based on precision (P), recall (R) and F1 score are shown for comparison and analysis.

	Precision			**Recall**			**F1 Score**		
	IE	IS	0	IE	IS	0	IE	IS	0
Multitask-attn-score	0.539	0.512	0.971	0.739	0.75	0.909	0.623	0.608	0.939
Multitask-score	0.614	0.485	0.938	0.712	0.68	0.887	0.660	0.566	0.912
Multitask	0.429	0.346	0.970	0.710	0.566	0.873	0.535	0.430	0.919
Multitask-attn	0.677	0.414	0.897	0.630	0.566	0.893	0.653	0.478	0.895

Table 6: Performance of the proposed models on Task B using the validation set. The traditional post-level evaluation metrics based on precision (P), recall (R) and F1 score are shown for comparison and analysis.

	Precision			**Recall**			**F1 Score**		
	Severe	Moderate	Low	Severe	Moderate	Low	Severe	Moderate	Low
Multitask-attn-score	0.555	0.500	0.00	0.625	0.357	0.00	0.588	0.416	0.00
Multitask-score	0.588	0.636	0.00	0.666	0.466	0.00	0.625	0.538	0.00
Multitask	0.764	0.300	0.00	0.565	0.428	0.00	0.650	0.352	0.00
Multitask-attn	0.846	0.400	0.000	0.523	0.666	0.00	0.647	0.500	0.000

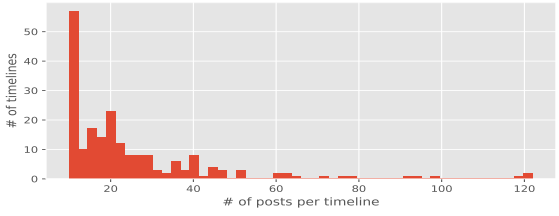

Figure 2: Distribution of number of posts per timeline in the training data set. The maximum number of posts in a timeline is 122.