Deliverable 1

Sofia Y. Ahmed
syahmed@uw.edu

Libbey Brown
egollhof@uw.edu

Rachel Hantz
hantz1rk@uw.edu

Elizabeth Okada
esokada@uw.edu

Abstract

This initial project report provides information about the tasks we will work on this quarter. Our primary task is multi-class emotion classification in English essays. Our adaptation task is multi-class emotion classification in Urdu tweets.

1 Introduction

2 Task description

2.1 Primary Task

Our primary task is emotion classification for essay texts written in English; this task is Track 2 of the shared task for WASSA 2022 (Barriere et al., 2022). We will develop a multi-class classification model which predicts an emotion tag from one of seven categories: Ekman's six basic emotions (joy, sadness, surprise, disgust, anger, or fear) and the neutral tag no-emotion.

The dataset was initially generated by Buechel et al. (2018) and consists of 2655 essays which were written by 403 participants in reaction to reading disturbing news articles. The dataset was downloaded from the WASSA (2022) website. Table 1 describes the train, development, and test splits.

Train	Dev	Test	Total
1860	270	525	2655

Table 1: Dataset Split for Primary Task

Participants rated their level of empathy and distress after reading the article and then described their thoughts and feelings in writing. Emotional tags were added to the essays as part of the WASSA 2021 shared task (Tafreshi et al., 2021). Emotion tags were added through prediction models (the specific models used were a gated RNN and a fine-tuned RoBERTA model) and the tags were manually verified by annotators from Amazon Mechanical Turk. Because these essays were generated in response to disturbing news articles, the distribution of emotion tags is not balanced; see Table 2 for the distribution of the emotion tags.

	Train	Dev	Test	Total
joy	82	14	33	129
sadness	647	98	177	922
disgust	149	12	28	189
fear	194	31	70	295
anger	349	76	122	547
surprise	164	14	40	218
no-emo	275	25	55	355

Table 2: Emotion Distribution for Primary Task

For evaluation, we will be using the evaluation script that was provided by the task organizers on the website page. This evaluation script reads in an input directory, which has two sub-directories. One sub-directory is "ref", which contains the gold standard file. The other sub-directory is the "res" directory, which contains .tsv file(s), one for each of the possible four tracks that task participants can choose from. We will be focusing on Track 2, emotion classification, so we will just have one .tsv file in the required format, which will be a single column with emotion predictions in string format. Our .tsv file is compared with the gold standard file and accuracy is calculated through several different measures: Macro F1-Score, Micro Recall, Micro Precision, Micro F1-Score, Macro Recall, Macro Precision, and Accuracy.

2.2 Adaptation Task

For our adaptation task, we plan to classify emotion for tweets in Urdu; this was Task A for the shared task at FIRE 2022 (Butt et al., 2023). As with our primary task, we plan to develop a multi-class classification model which predicts an emotion tag from one of seven categories: Ekman's six basic emotions (joy[1], sadness, surprise, disgust, anger, or fear) and the neutral tag no-emotion. The adaptation differs from our primary task in both language and genre.

The Urdu dataset consists of 9750 annotated tweets. The datatset was downloaded from the FIRE (2022) shared task website. Table 3 lists the split between training and test data. To mirror the primary task, we may create a development set extracted from the training data prior to model development.

Train	Test	Total
7800	1950	9750

Table 3: Dataset Split for Adaptation

The tweets are labeled with emotional tags among Ekman's six basic emotions and the neutral no-emotion tag. Construction of the dataset is detailed in Ashraf et al. (2022)[2]. The authors' goal was to construct a balanced dataset, so Twitter hashtags were used to identify tweets representative of a particular emotion, and then annotators classified the tweets with one or more emotion tags. Note that this dataset differs from the dataset used in our primary task in that each tweet can be tagged with more than one emotion (i.e. suggestive of a multi-label classification task). We will utilize only the tweets which have a single emotion tagged for ease of adaptation. Table 4 shows the distribution of emotion tags for the full Urdu data, prior to selecting only tweets with one emotion tag[3].

Table 5 lists the split between training and test data after selecting only tweets with one emotion tag.

	Train	Test	Total
joy	1046	261	1307
sadness	1550	388	1938
disgust	761	190	951
fear	609	152	761
anger	811	203	1014
surprise	1550	388	1938
no-emo	3014	753	3767

Table 4: Emotion Distribution for Adaptation

Train	Test	Total
5463	1389	6852

Table 5: Dataset Split for Adaptation: Only One Emotion Tag per Tweet

Table 6 shows the distribution of Urdu emotion tags after selecting only tweets with one emotion tag.

	Train	Test	Total
joy	744	185	929
sadness	911	230	1141
disgust	20	10	30
fear	159	41	200
anger	89	35	124
surprise	526	135	661
no-emo	3014	753	3767

Table 6: Emotion Distribution for Adaptation: Only One Emotion Tag per Tweet

We will use the same evaluation script as for the primary task.

[1]The original Urdu dataset uses 'happiness' in place of 'joy'. We map the label 'happiness' to 'joy.'

[2]The FIRE 2022 shared task references Ashraf et al. (2022) as a source for their dataset, but the present dataset appears to be a slightly larger sample than the one depicted in Ashraf et al. (2022). Thus, we display counts from the present dataset, not those described by Ashraf et al. (2022).

[3]285 train and 61 test tweets had no emotion tag.

3 System Overview

4 Approach

5 Results

6 Discussion

7 Ethical Considerations

8 Conclusion

References

Noman Ashraf, Lal Khan, Sabur Butt, Hsien-Tsung Chang, Grigori Sidorov, and Alexander Gelbukh. 2022. Multi-label emotion classification of urdu tweets. *PeerJ Computer Science*, 8:e896.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Sabur Butt, Maaz Amjad, Fazlourrahman Balouchzahi, Noman Ashraf, Rajesh Sharma, Grigori Sidorov, and Alexander Gelbukh. 2023. Emothreat@fire2022: Shared track on emotions and threat detection in urdu. New York, NY, USA. Association for Computing Machinery.

FIRE. 2022. Emothreat at fire 2022.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

WASSA. 2022. Wassa 2022 shared task.