Emotion Classification in English Essays and Urdu Tweets

Team Mood Masters:

Sofia Y. Ahmed

Libbey Brown

Rachel Hantz

Elizabeth Okada

Task Descriptions

Primary Task

- Emotion classification for English essay texts
  - Track 2 of the WASSA 2022 shared task (Barriere et al 2022)
- Multi-class classification system which predicts a single emotion tag from 7 options (Ekman's 6 basic emotions + neutral)

Adaptation Task

- Emotion classification for Urdu tweets
  - Task A of of EmoThreat FIRE 2022 shared task
- This differs in both genre and language
- The original version of this task was a multi-label task; we utilized only the tweets with a single tag so we only needed to adapt in genre and language!

Datasets

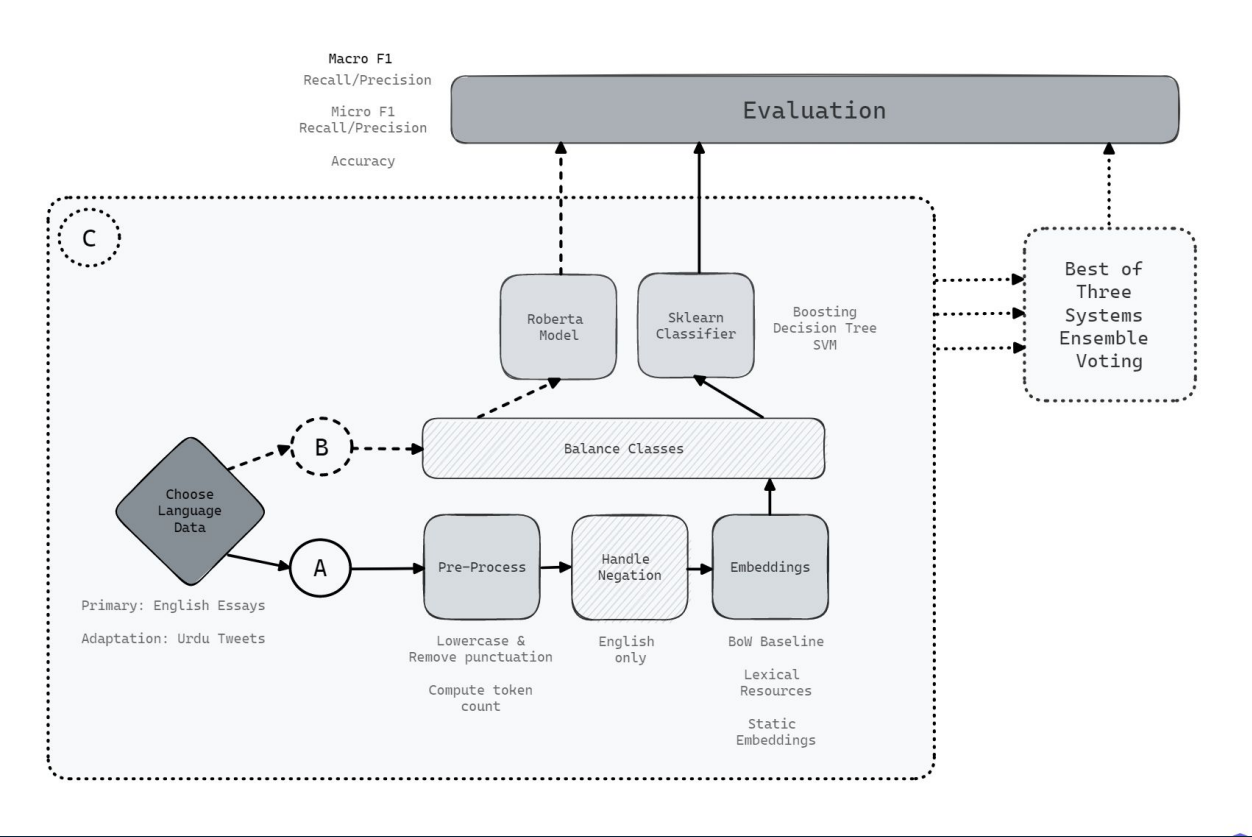
Primary

- Essay reactions to “disturbing” news articles (Buechel et al 2018)
  - Participants rated empathy and distress, emotion tags were added later
- 2700 essays tagged for single emotion
- Dataset is not balanced
  - ~35% are sadness, followed by anger
  - Joy is only ~3%

Adaptation

- Tweets selected based on hashtags indicative of a particular emotion, tagged by annotators (Ashraf et al 2022)
- Original dataset consisted of 9700 tweets tagged for zero through 6 emotions
- Our subset was 6800 tweets tagged for only one emotion
- Dataset is not balanced
  - ~55% are neutral, followed by sadness
  - Disgust is only ~1%

System overview



Core Approaches

A. Traditional machine learning models

- BoW, Emotion-Enhanced BoW vectors, Static Embeddings (e.g. w2v)
- Class balancing strategies
- Decision Tree classifier

B. Fine-tuned RoBERTa

- Primary: distilroberta-base
- Adaptation: urduhack/roberta-urdu-small

C. Ensemble Voting (*primary task only*)

- Best ⅔ voting for results of 3 best D3 systems

    1. balanced fine-tuned RoBERTa
    2. fine-tuned RoBERTA
    3. balanced emotion-enhanced BoW vectors + DT classifier

System Timeline (Dev)

Primary

D2 Best System: Macro F1: 0.2436
BoW vectors with DT classifier

→ Try negation, boosting, class balancing, RoBERTa

D3 Best System: Macro F1: 0.4588
RoBERTa with random over/under sampling each class to the average class size

→ Try Ensemble Voting

D4 Best System: Macro F1: 0.5026
Ensemble using unbalanced RoBERTa as tie-breaker

Adaptation

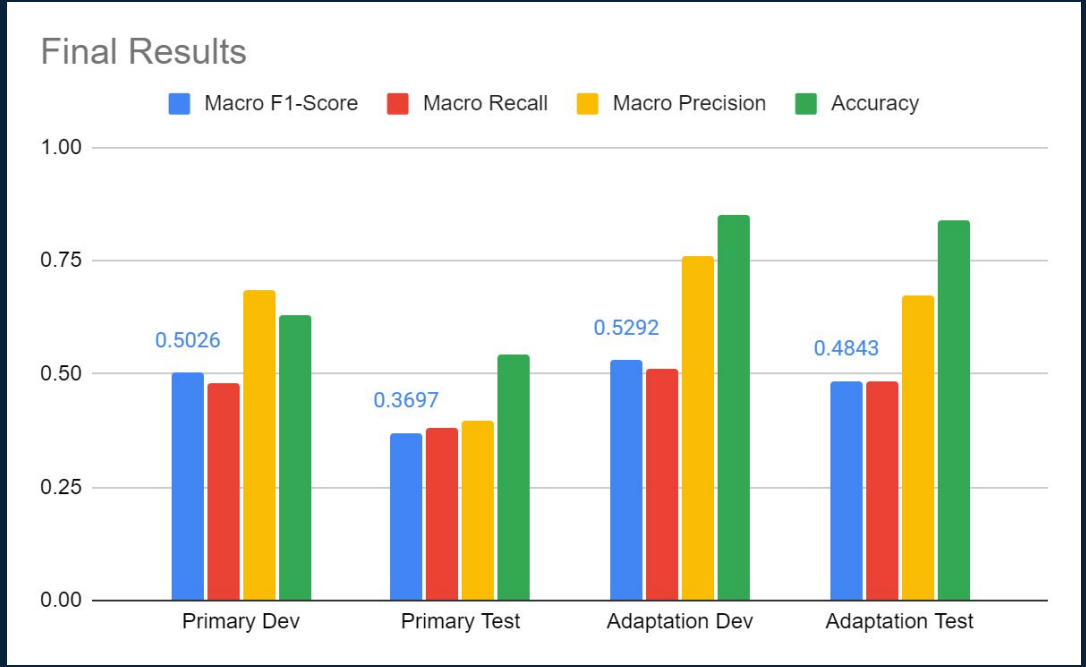
→ Try classic machine learning approaches

→ Try fine-tuning balanced and unbalanced RoBERTa models

D4 Best System: Macro F1: 0.5292
Unbalanced RoBERTa

Results

	Primary		Adaptation	
	Dev	**Test**	**Dev**	**Test**
Mac F1	**0.50**	**0.37**	**0.53**	**0.49**
Mac R	0.48	0.38	0.52	0.49
Mac P	0.69	0.40	0.76	0.67
Acc	0.63	0.55	0.85	0.84



* Accuracy = Micro Recall/Precision/F1

Issues and Successes

Primary

- Issues:
  - Class imbalance
  - Difficult to improve upon the baseline bag-of-words at first
  - Incorrect predictions falling into four categories:
    - Multiple emotions described
    - Nuanced expression of emotion
    - Minority class emotions
    - Poor annotation
  - Unable to perform data augmentation to compensate for the small size of training data
- Successes:
  - RoBERTa
  - Handling negation
  - Class balancing by random over/undersampling
  - Ensemble voting

Adaptation

- Issues:
  - The baseline was very hard to improve upon
  - Had to manually input proper label names in config file because when training the model it would encode the labels as LABEL_0, LABEL_1, etc.
  - Target and expected tensor shapes; had to set max length and specify truncation=true
- Successes
  - An easier task, perhaps because of more data, so baseline bag of words performed better than the same baseline in English.
  - First time training a model in a language other than English and it was fun!

Related Reading

- 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.
- Leung, K. (2022, September 13). Micro, Macro & Weighted Averages of F1 Score, Clearly Explained. Medium. https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f
- Batuwita, R., & Palade, V. (2013). Class Imbalance Learning Methods for Support Vector Machines. In Imbalanced Learning (pp. 83–99). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118646106.ch5
  Saif Mohammad and Peter Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.
- Saif M. Mohammad and Peter D. Turney. 2013. Crowd-sourcing a word-emotion association lexicon. Computational Intelligence, 29(3):436–465.
- 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.
- 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
- Chawla, N.V. (2009). Data Mining for Imbalanced Datasets: An Overview. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09823-4_45

Extras - Qualitative Error Analysis (Primary)

Correct predictions - author is straightforward and names their emotion

- My feelings after reading the article is that it made me feel a bit sad. (sadness)
- After reading the article, i was definitely a bit worried and alarmed. (fear)

Extras - Qualitative Error Analysis (Primary) - Incorrect Predictions

- Poor annotation:
  - *My feelings towards this article is neutral as well. It did not really make me feel anything. Rather yet, it made me feel a bit bored and i kind of dozed off for a bit. I felt like it wasn't anything that would impact me or anything i would think twice about. It made me feel mostly relaxed. I didn't have nay sort of positive or negative emotions towards the article. (gold = sadness, our system = neutral)*
- Least frequent emotion (joy):
  - *I am glad that they found the missing hiker and it is heartwarming that everything turned out for the best and he was found alive. It was truly amazing that he survived all that time without food or shelter. It goes to show that you should go looking for somebody missing no matter how long it takes because they might still be out there. It is an amazing story that will inspire and bring hope to the world (gold = joy, our system = sadness)*

Extras - Qualitative Error Analysis (Primary) - Incorrect Predictions

- More than one emotion:
  - *”I just read an artivcle about a police shooting. It was different than I expected. Two police officers where shot to death while sitting in their patrol cars. I have gotten so used to reading about them shooting and killing people I was surprised to read about them getting the same treatment. Everyday I am reading something or seeing a video of police shooting people. I am scared to death of getting pulled over. All the police in my area are rookies and do not know me from when I was a ””bad kid”” lol. I would be worried to death getting pulled over by the police seeing how trigger happy they have been lately.” (gold = surprise, our system = fear)*
- Nuanced description:
  - *Can you believe the Canadian government did something some terrible. I would never expect Canada to have something so bad on there record, especially given there reputation. Those families deserve justice and any living relatives deserve reparations from the government. This will at least be some justice (gold = disgust, our system = anger)*