Emotion Classification in English Essays

Team Mood Masters:

Sofia Y. Ahmed

Libbey Brown

Rachel Hantz

Elizabeth Okada

Task Description

- 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)
    - Joy
    - Sadness
    - Surprise
    - Disgust
    - Anger
    - Fear
    - Neutral (No Emotion)

Dataset

Essay reactions to "disturbing" news articles (Buechel et al 2018)

"It's a shame that air pollution has potentially been linked to increased mental damage with young children. We often don't take into account all the damage that the fossil fuel companies have done to our society. We only praise them for creating the fuels we use but never tax them appropriately for all the damage that they cause us."

[ANGER]

- Test data not available → Extract 10% from training data

  77.3% / 12.7% / 10% split

- Dataset is not balanced

  Sadness is ~35% of labels

Train	Dev	Test	Total
1860	270	525	2655

Table 1: Original Dataset Split for Primary Task

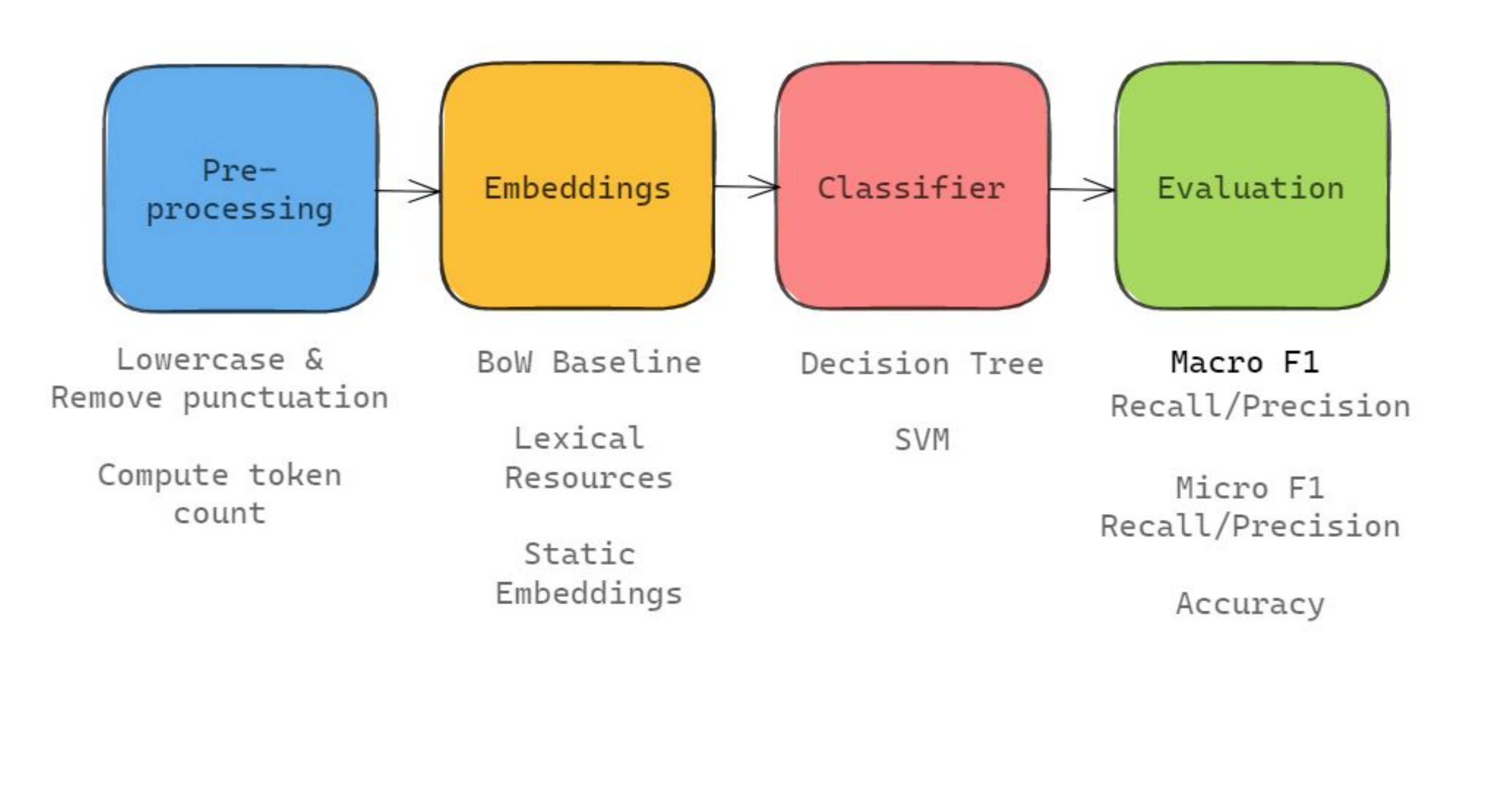
Train	Dev	Test	Total
1647	270	213	2130

Table 2: New Dataset Split for Primary Task

	Train	Dev	Test	Total
joy	72	14	10	96
sadness	570	98	77	745
disgust	131	12	18	161
fear	173	31	21	225
anger	312	76	37	425
surprise	145	14	19	178
no-emo	244	25	31	300

Table 3: Emotion Distribution for Primary Task

System overview



Embedding Approaches we Tried

- Baseline - Bag of Words
- Lexical Resources - NRC EmoLex

    Emotion Lexicon only vector
    Emotion Lexicon enhanced BoW vector

- Static Embeddings

    Word2vec trained on our training data
    Pre-trained Word2vec Google news corpus
    Pre-trained GloVe Twitter corpus

Lexical Resources
NRC EmoLex

(Mohammad and Turney, 2010, 2013)

	anger	disgust	fear	joy	sadness	surprise
shame	0	1	1	0	1	0
air	0	0	0	0	0	0
pollution	0	1	0	0	0	0

Example: "*...shame that air pollution...*"

Emotion Lexicon Only

[0, 2, 1, 0, 1, 0]

Emotion Lexicon Enhanced BoW - (concatenation of Bow + EmoLex Only)

[BoW] + [0, 2, 1, 0, 1, 0]

Static Embeddings

- Two versions of pre-trained static embeddings, trained on different domains
  - Word2Vec Google News - vectors trained on 100 billion word corpus from Google News ([Mikolov et al 2013](#))
  - GloVe Twitter - vectors trained on 27 billion tokens from 2 billion tweets ([Pennington et al 2014](#))
- Word2Vec embeddings created by training Word2Vec on our corpus of essays
- The centroid of each essay was calculated as the average of the vector representations of all words in the essay

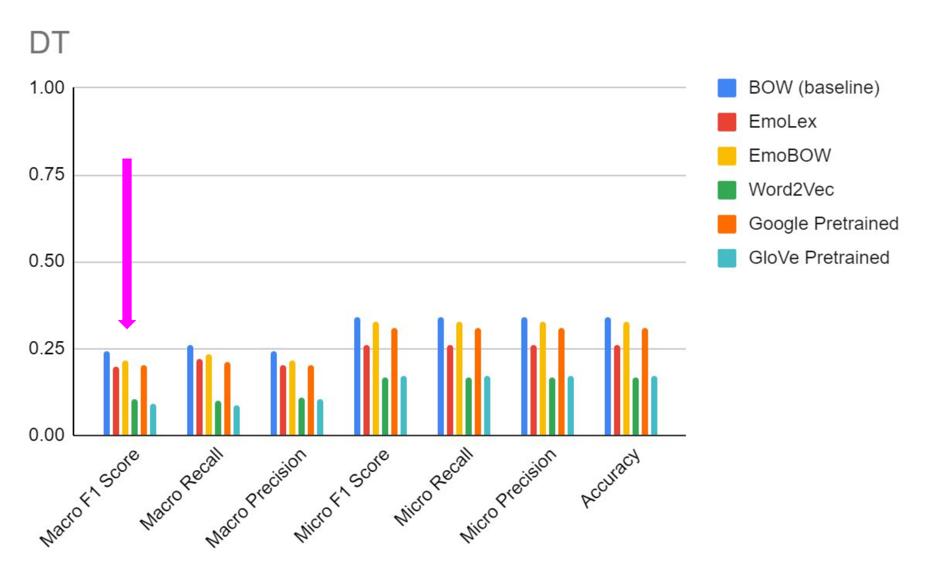
Classifiers

- SVM - yielded all sadness predictions, even when used together with various pre-trained embeddings
- DecisionTree - yielded a wider variety of predictions, but was not necessarily accurate

	A	D	F	J	N	Sa	Su
A	<u>24</u>	9	3	4	14	15	7
D	3	<u>5</u>	1	0	1	1	1
F	3	3	<u>9</u>	2	6	7	1
J	<u>6</u>	0	2	1	1	3	1
N	6	1	3	2	2	<u>8</u>	3
Sa	13	7	10	4	12	<u>49</u>	3
Su	<u>4</u>	1	3	1	1	2	2

Table 10: Confusion matrix displaying the number of predictions made on the BoW baseline using decision trees on the development data. The vertical axis indicates the ground truth emotion in alphabetical order: Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise. The horizontal axis indicates the predicted emotion. The highest predicted emotion for each class is <u>underlined</u>.

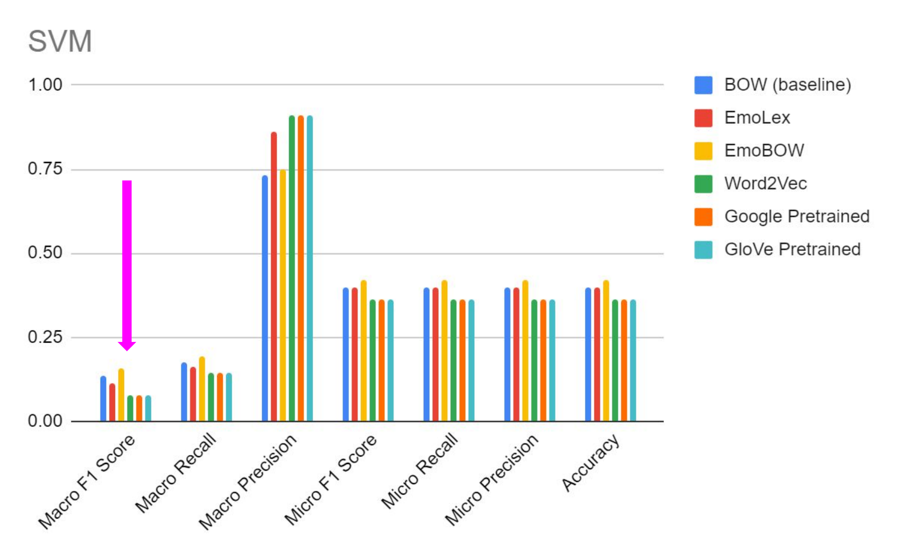
Error Analysis



Decision Tree is an objectively low performer, but has the highest macro F1: BoW baseline is the strongest.

SVM tends to over-predict the majority class, sadness.

AKA many true positives for sadness, but many false negatives elsewhere.



Best results

Macro F1: official competition metric for the shared task

Best performer: **Decision Tree** with baseline **BoW embeddings**

Next best performer: **Decision Tree** with **emotion lexicon enhanced BoW embeddings**

Theories as to why…

Vector	**Mac F1**	Mac R	Mac P	Mic F1
BOW	**24.36**	**25.96**	**24.32**	**34.07**
Emo	19.93	22.12	20.03	25.93
EmoBOW	21.56	23.52	21.54	32.59
W2V	10.22	10.09	11.03	16.67
W2V PT	20.34	21.13	20.23	30.74
GloVe PT	9.16	8.8	10.42	17.04

Table 9: Macro F1-score, macro recall, macro precision, and micro F1-score using *decision tree* classifier

Limitations & Future Directions

D3

- **Pre-processing:** Handle class imbalance and negation
- **Embeddings:** Try contextualized embeddings
- **Classification:** Neural classifiers, Ensemble learning
- **Evaluation:** Examine weighted and per-class metrics

D4

- **Adaptation:** Genre & Language (Tweets in Urdu)

Things We Learned

- Don’t try downloading ELMo embeddings without compute!
- Test your conda environment early and often
- No one evaluation metric tells the whole story
- Code provided by task organizers can be hard to use

References

- 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.

- Gennaro, Gloria, and Elliott Ash. "Emotion and reason in political language." *The Economic Journal* 132.643 (2022): 1037-1059.

- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013) *Advances in neural information processing systems* 26 (2013).

- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

- 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.