VISU at WASSA 2023 Shared Task: Detecting Emotions in Reaction to News Stories Using Transformers and Stacked Embeddings

Vivek Kumar
University of Cagliari, Italy
vivek.kumar@unica.it

Prayag Tiwari
Halmstad University, Sweden
prayag.tiwari@hh.se

Sushmita Singh
Liverpool John Moores University, UK
sushmitafordata@gmail.com

Abstract

Our system, VISU, participated in the WASSA 2023 Shared Task (3) of Emotion Classification from essays written in reaction to news articles. Emotion detection from complex dialogues is challenging and often requires context/domain understanding. Therefore in this research, we have focused on developing deep learning (DL) models using the combination of word embedding representations with tailored prepossessing strategies to capture the nuances of emotions expressed. Our experiments used static and contextual embeddings (individual and stacked) with Bidirectional Long short-term memory (BiLSTM) and Transformer based models. We occupied rank tenth in the emotion detection task by scoring a Macro F1-Score of 0.2717, validating the efficacy of our implemented approaches for small and imbalanced datasets with mixed categories of target emotions.

1 Introduction

Digitalization and ease of access to internet-based intelligent and interactive technologies have led to an unprecedented amount of textual data generation from social media, customer reviews, and online forums. Therefore, the need to accurately understand and extract emotions and sentiments from text has become imperative for two reasons; first, due to their various crucial applications such as sentiment analysis (Gupta et al., 2023), chatbots, mental health assessment(Wu et al., 2020), social media monitoring, market research, brand management, and customer feedback analysis and second to reduce the human efforts, time and resource requirements. The Shared Task on *Empathy Detection, Emotion Classification and Personality Detection in Interactions* of WASSA 2023[1] aims to develop models to predict various targets, including emotion, empathy, personality, and interpersonal-index, from textual data (Barriere et al., 2023). The shared task consists of five tracks, of which we participated in *Track 3: Emotion Classification* (EMO), which targets emotion classification at the essay level. This work presents two systems to capture the subtle notion of emotions expressed through texts: a) BiLSTM-based (Graves and Schmidhuber, 2005) DL model using static, contextual, and the combination of static and contextual (stacked) embeddings and b) Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Stacked embeddings (Bhandari et al., 2022) are fast-to-train, powerful but underutilized representations; therefore, to reckon their efficacy compared to the transformer model, we have used them in this work. Our proposed systems have performed competitively and got the tenth rank[2] in the evaluation phase of the *Track 3: EMO* task.

The remainder of the paper is structured as follows: Section 2 presents the notable research works on emotion detection. Section 3 presents the problem statement, dataset description, and the preprocessing strategies applied. In section 4, we present our different classification systems and the experimental setup. Section 5 presents the evaluation results of our proposed systems and comparison with other participating teams of the shared task. Finally, section 6 provides the conclusion and discusses the future research directions.

2 Literature Survey

The significance of accurate emotion detection and sentiment analysis extends beyond understanding textual data. Recent research has brought the machines one step closer to mimicking humans' innate ability to understand emotional cues from and text and different modalities. Works such as (Acheampong et al., 2020; Chatterjee et al., 2019) explore the emotion detection form texts; (Zhong

[1] https://2023.aclweb.org/program/workshops/

[2] The rank is solely based on the submissions done before the deadline of the shared task

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 581–586
July 14, 2023

et al., 2019; Acheampong et al., 2021; Adoma et al., 2020) explores the variants of transformer models useful for emotion detection from texts. Some notable works such as (Wu et al., 2023, 2022; Bostan et al., 2020; Bostan and Klinger, 2018; Buechel and Hahn, 2017; Sosea and Caragea, 2020) have created the novel datasets from textual and conversational settings to address the scarce data challenges in complex domains for emotion detection.

3 Problem Statement, Dataset Description and Data Preprocessing

In this section, we have mentioned the problem statement tackled, the dataset description, and the data-prepossessing techniques implemented for our experiments.

3.1 Problem Statement

In this work, we tackled a multiclass classification problem to predict emotions from essay-level texts. The target labels consist of thirty-one categories of emotions, including individual and mixed sets of emotion categories, as follows: Hope/Sadness, Anger, Sadness, Neutral, Disgust/Sadness, Anger/Disgust, Fear/Sadness, Joy, Hope, Joy/Neutral, Disgust, Neutral/Sadness, Neutral/Surprise, Anger/Neutral, Hope/Neutral, Surprise, Anger/Sadness, Fear, Anger/Joy, Disgust/Fear, Fear/Neutral, Fear/Hope, Joy/Sadness, Anger/Disgust/Sadness, Anger/Surprise, Disgust/Neutral, Anger/Fear, Sadness/Surprise, Disgust/Surprise, Anger/Hope, and Disgust/Hope.

3.2 Dataset Description

The experimental dataset contains long essays of length between 300 and 800 (Omitaomu et al., 2022). The dataset includes news articles and person-level demographic information (empathy, distress, age, race, income, gender, education level, emotion labels, etc.). The dataset was made available as training, development (dev), and test sets where the target labels were shared only for the training and development sets for the evaluation phase. The overall distribution of the dataset is shown in Table 1 and the distribution of each emotion class of the train and dev sets is shown in Table 1.

Dataset Split Distribution			
Train	*Dev*	*Test*	*Total*
792	208	100	1000

Table 1: Train, dev, and test set distribution.

3.3 Dataset Preprocessing

As evident from Figure 1, the dataset is small and imbalanced and several emotion categories have only one data point. Also, the mixed categories of emotions in the target class made the task more challenging. Therefore, to overcome these constraints, we have applied a tailored preprocessing strategy along with standard NLP techniques to prepare the input dataset (Dessì et al., 2020; Kumar et al., 2021; Uysal and Gunal, 2014). The preprocessing steps are as follows. The input texts are converted to lowercase to make the dataset uniform in terms of representation (e.g., Emotion and emotion are represented by a common token, emotion). Punctuation, stopwords, newlines, whitespaces, and extra spaces are removed from the text. We have removed the special characters, symbols, and elements which are not part of the standard English language. We have expanded the contractions such as *didn't –> did not*. We performed stemming and lemmatization alternatively for experiments but observed a slight decline in the model's performance. Therefore, we have not considered them for preprocessing the input dataset for the final submission of *Track 3: EMO* shared task.

4 Methodology

This section describes our different systems (classification models) based on the BiLSTM and transformer model implemented for the emotion classification task.

4.1 BiLSTM Based DL Model

Our first system is a DL model using two BiLSTM layers. More precisely, this model's architecture consists of an embedding layer, followed by two BiLSTM layers, a dense layer, and an output layer at the end for the multi-class classification. The embedding layer is initialized by input_dim (size of the vocabulary); output_dim: (word vector length), embedding matrix, and sequences length. For ease of understanding, we have summed up the parameters and combination of embeddings used for our experiments in Table 2.

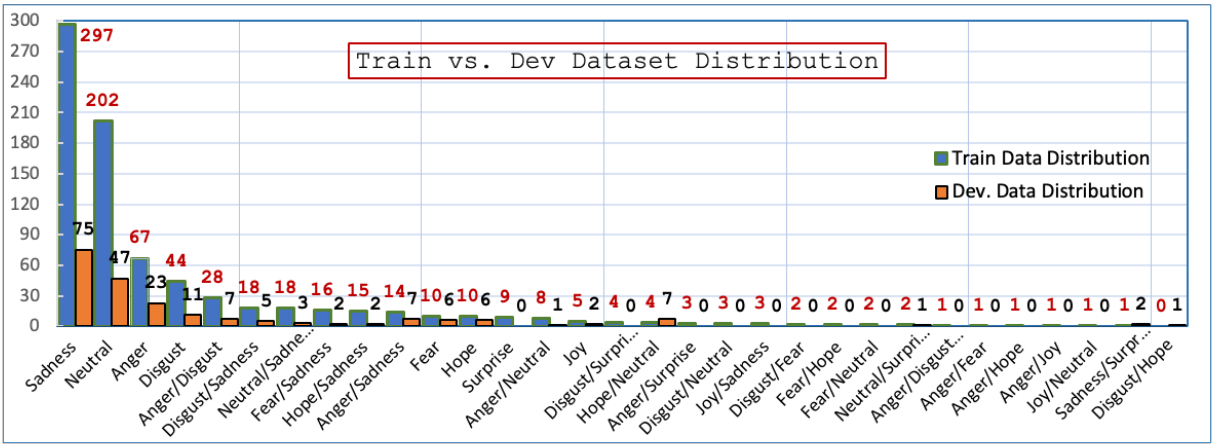

Figure 1: Plot showing the skewed distribution of training and development dataset.

4.2 BERT

The second system is a transformer-based model *BERT*, created using *Keras*[3] and Tensorflow[4]. Our *BERT* model comprises two input layers, a *BERT* model layer, and two dense layers of 768 embedding dimensions with the Adam optimizer. The parameters used for fine-tuning the model are listed in Table 2.

4.3 Features Representation

We have used pre-trained static and contextual word embeddings for our experiments to generate the feature vectors discussed below.

GloVE (Global Vectors for Word Representation): *GloVE*[5] is an unsupervised learning algorithm that generates word embeddings as dense vector representations of words in a high-dimensional space. It leverages co-occurrence statistics from a large text corpus to capture semantic relationships between words. *GloVe* embeddings are trained by factorizing a matrix representing the word co-occurrence statistics (Pennington et al., 2014).

fastText: *fastText*[6] was developed by Facebook's AI Research (FAIR) team (Bojanowski et al., 2017; Joulin et al., 2016). *fastText* extends the traditional word embeddings by representing each word as a bag of character n-grams, where n can range from 1 to a maximum specified length. This approach allows *fastText* to capture morphological information and handle out-of-vocabulary words effectively.

BERT: *BERT*[7] embeddings are a type of word representation that captures contextual information in the text. Unlike traditional word embeddings like *Word2Vec* or *GloVe*, *BERT* embeddings take into account the surrounding words when representing a word. This means that the meaning of a word can vary depending on its context.

FLAIR (FastText and Language-Independent Representations): *FLAIR*[8] embedding is a state-of-the-art word representation model that captures contextual information and word semantics by combining the strengths of two powerful techniques: *FastText* and contextual string embeddings. By combining these techniques, *FLAIR* embedding provides a robust and language-independent representation of words. It considers both the local context of a word and its global context within a sentence or document (Akbik et al., 2019).

5 Results and Analysis

Table 3 presents the results of our two systems. The BERT base system has significantly outperformed the BiLSTM-based system using the combination of *GloVe*, *fastText* and *BERT* embeddings. Therefore, we have submitted the BERT base result for the shared task evaluation phase. The evaluation of *Track 3: EMO* shared task is based on the macro F1-score and Micro Jaccard-score, Micro F1-score, Micro Precision, Micro Recall, Macro Precision and Macro Recall are supporting metrics. Our BERT base system has achieved a Macro F1-score of 2.717 and stood tenth[9] among participants. Table 4 presents the official results of all the qualifying teams.

[3]https://keras.io/
[4]https://t fhub.dev/google/collections/bert
[5]https://nlp.stanford.edu/projects/glove/
[6]https://fasttext.cc/
[7]https://pypi.org/project/bert-embedding/
[8]https://pypi.org/project/flair/
[9]These rankings were provided by the shared task organizing team on 11/05/2023.

Models	*Embedding Dimension*	*Sequence Length*	*Batch Size*	*Epoch*	*Learning Rate*
BiLSTM + GloVe	100	74	32	3	0.001
BiLSTM + fastText	300	74	32	4	0.001
BiLSTM + (GloVe & fastText)	400	128	32	3	0.001
BiLSTM + (GloVe & BERT)	868	128	32	3	0.001
BiLSTM + (fastText & BERT)	1068	152	32	7	0.001
BiLSTM + (GloVe, fastText & BERT)	1168	152	32	5	0.001
BERT	768	152	32	5	2e-5

Table 2: Experimental settings of proposed systems.

DL Model	**Static Embedding**		**Stacked Embedding**				**Contextual**
Metrics	*GloVe*	*fastText*	*GloVe + fastText*	*GloVe + BERT*	*fastText + BERT*	*GloVe+ fastText +BERT*	*BERT*
Micro F1-Score	0.213	0.213	0.213	0.213	0.204	0.213	**0.593**
Macro F1-Score	0.075	0.075	0.075	0.075	0.073	0.075	**0.284**
Micro Jaccard	0.119	0.119	0.119	0.119	0.113	0.119	**0.421**
Micro Precision	0.230	0.230	0.230	0.230	0.220	0.230	**0.640**
Macro Precision	0.046	0.046	0.046	0.046	0.045	0.046	**0.282**
Micro Recall	0.198	0.198	0.198	0.198	0.19	0.198	**0.552**
Macro Recall	0.192	0.192	0.192	0.192	0.183	0.192	**0.318**

Table 3: The results of our implemented models for static and contextual embeddings.

Rank	**Team ID**	**Macro F1 Score**	**Micro Recall**	**Micro Precision**	**Micro F1 Score**	**Macro Recall**	**Macro Precision**	**Micro Jaccard**
1	adityapatkar	0.7012	0.7241	0.7778	0.750	0.6773	0.8105	0.600
2	anedilko	0.6469	0.7931	0.6259	0.6996	0.7305	0.6305	0.538
3	luxinxyz	0.644	0.6983	0.7431	0.72	0.6314	0.7207	0.5625
4	zex	0.6426	0.7069	0.7321	0.7193	0.637	0.6992	0.5616
5	lazyboy.blk	0.6125	0.6638	0.77	0.713	0.6005	0.7764	0.554
6	gauravk	0.5649	0.7069	0.6949	0.7009	0.5605	0.5955	0.5395
7	amsqr	0.533	0.6293	0.7228	0.6728	0.4793	0.7521	0.5069
8	surajtc	0.522	0.7586	0.5269	0.6219	0.6679	0.4626	0.4513
9	alili_wyk	0.5142	0.6724	0.7358	0.7027	0.5022	0.575	0.5417
10	**kunwarv4**	0.2717	0.5517	0.64	0.5926	0.3012	0.2571	0.4211
11	Cordyceps	0.202	0.4138	0.3664	0.3887	0.2356	0.1905	0.2412
12	Sidpan	0.1497	0.4138	0.4848	0.4465	0.2111	0.2948	0.2874
13	mimmu3302	0.126	0.3966	0.46	0.4259	0.2	0.092	0.2706

Table 4: The official results of the evaluation phase of *Track 3: EMO* task. Our system VISU (Team ID kunwarv4 attained the tenth rank.)

6 Conclusion

Our system, VISU, participated in the shared task *Track 3: EMO* of emotion classification tasks of the WASSA 2023, and our BERT base system scored tenth rank. Our experiments conclude that although *FLAIR* are powerful word representations built to capture *out-of-vocabulary* words, they are not as effective as contextual embeddings when used for small and imbalanced datasets. Our future research aims to address the data imbalance and scarce data challenges (Kumar. et al., 2023) by incorporating novel augmentation techniques of domain adaptation(Kumar et al., 2022) to interpret better the emotions expressed in text.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, pages 1–41.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121. IEEE.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. Wassa 2023 shared task: Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Aakash Bhandari, Vivek Kumar, Pham Thi Thien Huong, and Dang NH Thanh. 2022. Sentiment analysis of covid-19 tweets: Leveraging stacked word embedding representation for identifying distinct classes within a sentiment. In *Artificial Intelligence in Data and Big Data Processing: Proceedings of ICABDE 2021*, pages 341–352. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Danilo Dessì, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. 2020. TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study. In *Proceedings of the First Workshop on Smart Personal Health Interfaces co-located with 25th International Conference on Intelligent User Interfaces, SmartPhil@IUI 2020, Cagliari, Italy, March 17, 2020*, volume 2596 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Shelley Gupta, Archana Singh, and Vivek Kumar. 2023. Emoji, text, and sentiment polarity detection using natural language processing. *Information*, 14(4).

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Vivek Kumar., Simone Balloccu., Zixiu Wu., Ehud Reiter., Rim Helaoui., Diego Recupero., and Daniele Riboni. 2023. Data augmentation for reliability and fairness in counselling quality classification. In *Proceedings of the 1st Workshop on Scarce Data in Artificial Intelligence for Healthcare - SDAIH,*, pages 23–28. INSTICC, SciTePress.

Vivek Kumar, Diego Reforgiato Recupero, Daniele Riboni, and Rim Helaoui. 2021. Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. *IEEE Access*, 9:7107–7126.

Vivek Kumar, Diego Reforgiato Recupero, Rim Helaoui, and Daniele Riboni. 2022. K-lm: Knowledge augmenting in language models within the scholarly domain. *IEEE Access*, 10:91802–91815.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Tiberiu Sosea and Cornelia Caragea. 2020. Canceremo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904.

Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. 2020. Towards detecting need for empathetic response in motivational interviewing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 497–502.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.