Human-centred training and validation of text-based emotion detection machine learning models

Zayn J. Abbas*

School of Computer Science, University of Guelph, Canada, abbasr@uoguelph.ca

Stacey D. Scott

School of Computer Science, University of Guelph, Canada, stacey.scott@uoguelph.ca

When communicating via social media, people rely heavily on text-based emotional expression. However, interpreting and understanding emotions expressed by others over text is challenging due to the lack of non-verbal communication cues such as tone and body language. Advances in Artificial Intelligence (AI), especially in natural language processing and large-language models, have made significant advances in nuanced language interpretation. Yet, human emotion is a complex psychological phenomenon. AI-based emotion detection is still an emerging field, lacking comparative studies and human-oriented design approaches. In this research, we take a human-centred approach to developing and validating emotion detection machine learning (ML) models in the context of text-based social media communication. We propose an interdisciplinary approach in which a natural language processing model, based on the popular RoBERTa ML model, is trained on datasets of social media communications labelled using different psychological emotion theories. The trained emotion detection models are then validated through a user study that compares the model output to human evaluators to determine which model best replicates human emotion interpretation.

Keywords and Phrases: Emotion Detection • Human-Computer Interaction • Artificial Intelligence

1 INTRODUCTION

Emotion detection and interpretation is a complex field due to the multidimensional factors of human language understanding. When recognizing human emotions through speech, individuals rely heavily on physical cues, such as body language, tone of voice, and facial expressions [7]. With the rise of social media, people rely heavily on emotional expression through posts, comments, and social media interactions [8]. Emotion interpretations vary across cultures, age groups, and societal groups [5,6]; thus, the lack of non-verbal communication cues on social media text posts adds another layer of complexity to emotion recognition.

Advances have been made within the field of sentiment analysis and emotion detection to develop and improve machine learning (ML) models for social media platforms [8,16,17]. However, many of these models cater to industry sectors, such as businesses that want to gather information on their products based on customer reviews or researchers looking into political discourse [1,2]. While some work has been done in terms of depression, hate speech, and sarcasm detection, there is a gap in the research regarding assisting users in emotion detection and interpretation [2,9,15].

* Place the footnote text for the author (if applicable) here.

The ultimate goal of this research is to develop reliable emotional detection models to help social media users interpret intended emotion of text-based posts. This may help minimize misunderstandings in social media-based communications, and also improve the accessibility of social media communications for people who have issues recognizing emotions, for instance, some autistic people [10]. An important first step towards this goal is to develop reliable emotion detection models. This research proposes a new methodology to develop a reliable model and to validate its effectiveness.

2 RESEARCH PROBLEM

The lack of non-verbal communication cues, anonymity, and other factors create challenges for accurately interpreting text-based communications on social media. Such misunderstandings can lead to interpersonal conflict, psychological stress, and social isolation. Developing ML tools for automated emotion detection in text-based communication platforms may help prevent such misunderstandings and result in more positive social outcomes. However, despite advancements in ML models for emotion detection, limitations and inconsistencies exist in data labeling and in model training and validation [1,2,12,13]. Prior research has found inconsistencies in emotion labels due to discrepancies in their meanings [14,16]. Also, many emotion detection ML models are rarely validated against human interpretations, raising questions about their applicability in real-world settings [2,8,16].

3 PROPOSED SOLUTION

To tackle this problem, we take an interdisciplinary approach at the intersection of AI, psychology, and human-computer interaction to introduce a novel methodology for emotion detection ML training and validation. The proposed methodology is divided into two stages, as described below.

Stage 1: Human-centred model training. In this stage we will create multiple emotion detection ML models. To create these emotion detectors, we will first create custom datasets for training and model performance testing. Each dataset will use the same underlying text-based content derived from social media posts. Each version of the dataset will be labeled using a different prominent theoretical emotion model from the psychology literature, for instance, the Categorical theory of emotion and the Dimensional theory of emotion [3,4,7,11]. Each emotion model dataset will then be used to train different emotion detection ML models, using the underlying TweetNLP RoBERTa-based ML model. Data will be gathered from X (formerly Twitter), with a focus on LGBTIQA+ college and university students for this project.

Stage 2: Human-centred model validation. To validate the emotion detection ML models created in stage 1, we will compare the output against human evaluators of the same text-based dataset. An online user study will be conducted using a crowd work platform such as Amazon Mechanical Turk. Participants will be asked to interpret the intended emotion of text-based content based on a set of provided emotion labels (based on the respective models used for emotion training). Participants' emotional intelligence will also be measured using standardized psychological scales for this personality trait.

The human-based emotion interpretation will then be compared with the output of the different emotion detector ML models to determine the level of agreement between the ML models and human interpreters. The data will also be analyzed to determine whether human-based interpretation of the data and the level of agreement with the respective models are impacted by participants' emotional intelligence scores.

4 NEXT STEPS

We have created our initial models, and selected our emotion theories to use for stage 1 of the research. The immediate next step in the research will involve gathering the needed social media data for creating our training and performance testing datasets. Selection of an appropriate crowd work platform and conducting the user study will follow.

REFERENCES

1. Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports* 2, 7. https://doi.org/10.1002/eng2.12189

2. Mashary N. Alrasheedy, Ravie Chandren Muniyandi, and Fariza Fauzi. 2022. Text-Based Emotion Detection and Applications: A Literature Review. In *International Conference on Cyber Resilience, ICCR 2022*. https://doi.org/10.1109/ICCR56254.2022.9995902
3. Saima Aman and Stan Szpakowicz. 2007. *Identifying Expressions of Emotion in Text*. Retrieved from http://www.lrec-conf.org/lrec2006/IMG/pdf/programWSemotion-LREC2006-last1.pdf
4. Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. In *Computational Intelligence*, 527–543. https://doi.org/10.1111/j.1467-8640.2012.00456.x
5. Robert M Chapman, Margaret N Gardner, and Megan Lyons. 2022. Gender differences in emotional connotative meaning of words measured by Osgood's semantic differential techniques in young adults. *Humanities and Social Sciences Communications* 9, 1: 119. https://doi.org/10.1057/s41599-022-01126-3
6. Yingruo Fan, Jacqueline C K Lam, and Victor O K Li. 2021. Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness. *Scientific Reports* 11, 1: 5214. https://doi.org/10.1038/s41598-021-84632-9
7. Elaine Fox. 2008. *Emotion Science*. Macmillan Education UK, London. https://doi.org/10.1007/978-1-137-07946-6
8. Bharat Gaind, Varun Syal, and Sneha Padgalwar. 2018. Emotion Detection and Analysis on Social Media. *Global Journal of Engineering Science and Researches* ICRTCET-18: 78–89. Retrieved from http://arxiv.org/abs/1901.08458
9. Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Computing Surveys* 55, 14s: 1–35. https://doi.org/10.1145/3583067
10. Connor T Keating, Eri Ichijo, and Jennifer L Cook. 2023. Autistic adults exhibit highly precise representations of others' emotions but a reduced influence of emotion representations on emotion recognition accuracy. *Scientific Reports* 13, 1: 11875. https://doi.org/10.1038/s41598-023-39070-0
11. GS Mahalakshmi. 2017. *Emotion Models: A Review*. Retrieved from https://www.researchgate.net/publication/319173333
12. Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, M. F. Mridha, and Zeyar Aung. 2024. Challenges and Opportunities of Text-Based Emotion Detection: A Survey. *IEEE Access* 12: 18416–18450. https://doi.org/10.1109/ACCESS.2024.3356357
13. Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining 11*. https://doi.org/10.1007/s13278-021-00776-6
14. Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining* 8, 1. https://doi.org/10.1007/s13278-018-0505-2
15. Xuanyu Su, Yansong Li, Paula Branco, and Diana Inkpen. 2023. SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media. *Natural Language Engineering*. https://doi.org/10.1017/S1351324923000396
16. M Usman Ashraf, Moeed Rehman, Qasim Zahid, Mustahsan Hammad Naqvi, and Iqra Ilyas. *A Survey on Emotion Detection from Text in Social Media Platforms*.
17. Yuxi Wang, Diana Inkpen, and Prasadith Buddhitha. 2024. *Explainable Depression Detection Using Large Language Models on Social Media Data*.

HUMAN-CENTRED TRAINING AND VALIDATION OF TEXT-BASED EMOTION DETECTION MACHINE LEARNING MODELS

AUTHORS

affiliations

INTRODUCTION

Text-based emotion detection has multiple challenges, including detection, classification, and interpretation. In-person interactions, humans rely heavily on physical cues, such as body language, tone, and facial expression, to interpret emotions [4]. While many use social media to express their feelings and opinions, social media also causes isolation for those who struggle to understand these posts [5].

OBJECTIVE

This research project aims to develop improved approaches for automated emotion detection of text-based social media content. We aim to introduce a novel methodology for the emotion detection process. Starting from data gathering and labelling to model training, testing, and evaluation.

RESEARCH PROBLEM

Many advancements have been made in Natural Language Processing (NLP), specifically sentiment analysis (determining the polarity of a text) and emotion detection. However, there are inconsistencies in emotion labels and discrepancies in their meanings [1,2]. Additionally, many emotion detection models are not validated in real-world settings or by human evaluators [1,2].

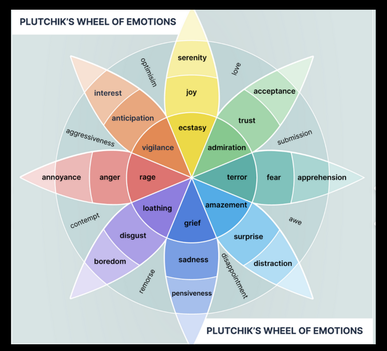
METHODOLOGY

The proposed methodology includes the following:

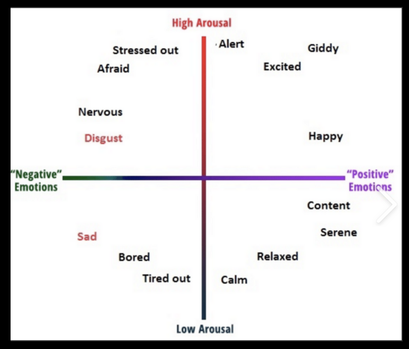
- **Identifying psychological emotion models**
- **Labeling data based on emotion models**
- **Training a machine learning model on the given data**
- **Have human evaluators label the test data**
- **Compare the results of the human evaluator data to that of the models**

PSYCHOLOGICAL EMOTION MODELS

CATEGORIAL THEORY OF EMOTION



DIMENSIONAL THEORY OF EMOTION



PILOT STUDY RESULTS

The pilot study was intended to understand better how humans interpret emotion and how the trained model compares the human evaluators. The goal of this study includes:

GOALS

- ***Training an ML model on a mix between Categorical & Dimensional emotion models [3]***
- ***Validate the results of the model with human evaluators***

FUTURE WORK

- Collect Twitter dataset based on LGBTIQA+ college/university students.
- Select psychological emotion models: one Categorical & one Dimensional model.
- Conduct the proposed method to train, test, and evaluate RoBERTA based ML model.
- Crowdsourcing human evaluators to select the most appropriate label for the test data.
- Conduct comparative analysis on the results of the human evaluators and model output

University and funding logos - removed for blind review

REFERENCES

1. Acheampong, F. A. et al. 2020. Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports 2, 7.
2. Alrasheedy, M.A. et al. 2022. Text-Based Emotion Detection and Applications: A Literature Review. In International Conference on Cyber Resilience, ICCR 2022.
3. Calvo, R.A. & Kim, S.M. 2013. Emotions in text: Dimensional and categorical models. In Computational Intelligence, 527–543.
4. Fox, E. 2008. Emotion Science. Macmillan Education UK, London.
5. Gaind, B. et al. 2018. Emotion Detection and Analysis on Social Media. Global Journal of Engineering Science and Researches ICRTCET-18: 78–89