Emotion Analysis from Facebook Messages
Project Proposal for NLP Course, Winter 2024

Adam Czerwoński
Warsaw University of Technology
adam.czerwonski.stud@pw.edu.pl

Jakub Kubacki
Warsaw University of Technology
jakub.kubacki.stud@pw.edu.pl

Jedrzej Ruciński
Warsaw University of Technology
jedrzej.rucinski.stud@pw.edu.pl

Maja Wasielewska
Warsaw University of Technology
maja.wasielewska.stud@pw.edu.pl

Supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

The proposed project aims to review available resources for emotion classification in text, focusing on private data such as personal Facebook messages. Despite significant advancements in Natural Language Processing (NLP), little research has explored its application to private datasets for psychological self-assessment.

The proposed deliverable is a comprehensive evaluation of NLP models for emotion classification on private data, highlighting their potential and challenges. We will compare Polish-specific NLP models with English models applied to translations of the original Polish text, assessing the impact of translation on the results.

1 Introduction

1.1 Scientific Goal of the Project

The scientific goal of this project is to evaluate the capabilities of emotion classification models in handling raw, unfiltered private data, with a focus on assessing their potential for providing psychological insights. Specifically, the project addresses three key research questions:

(1) How well do existing pre-trained emotion classification models perform on new, unstructured datasets such as private Facebook messages?

(2) What are the comparative results between Polish-specific models and English models applied to translations of Polish text?

(3) To what extent can emotion classification models support self-assessment in psychological contexts?

The central hypothesis is that existing NLP models can provide meaningful emotional insights when applied to private data but may exhibit limitations due to translation or lack of domain-specific fine-tuning.

1.2 Significance of the Project

This research is significant as it tackles the under-explored intersection of NLP, private data analysis, and psychology. While the state of the art in emotion classification has advanced significantly, most studies rely on publicly available or pre-curated datasets, which lack the complexity and authenticity of raw, personal communications. Moreover, the potential impact of translation on emotion classification in multilingual contexts remains largely unstudied. By introducing private Facebook messages as a dataset and examining both Polish-specific and English-translated models, this project breaks new ground in evaluating how language and context influence model performance.

The results could provide practical guidance for using emotion classification tools in personal and psychological contexts, such as improving tools for mental health monitoring or self-awareness. This work aims to bridge the gap between theoretical model performance and their application to real-world, sensitive data.

1.3 Literature review

Emotion classification in text is a part of Natural Language Processing (NLP) focused on figuring out and labeling the emotions people express in their writing. It's used in things like sentiment analysis, understanding customer opinions, analyzing feedback, or keeping track of what's trending on social media. This field intersects with psychology, computational linguistics, and machine learning.

The foundational theories for emotion classification often derive from psychological models. Emotions can be described discretely, for instance, as one of the six basic emotions proposed by Paul Ekman (Ekman, 1992). Alternatively, dimensional models evaluate emotions along various dimensions, such as arousal and valence (Russel, 1980).

Early classification works relied on rule-based systems that utilized pre-defined lexicons and linguistic rules (Pennebaker et al., 1999; Strapparava et al., 2004; Mohammad et al., 2013). Those solutions were transparent and simple however didn't generalize well to domain-specific language and evolving textual expressions.

With advancements in computational power, machine learning methods became popular (Roberts et al., 2012; CBalabantaray et al., 2012; Hasan et al., 2014). Models such as Support Vector Machines (SVMs), Naïve Bayes, and Random Forests trained on feature-engineered datasets (e.g., TF-IDF, n-grams, POS tags) were used. However, this methods were very dependent on feature engineering and struggled with capturing contextual nuances.

The recent wave of advancements in emotion classification has been driven by deep learning methods, which have transformed the field. Early approaches like RNNs (including LSTMs and GRUs) and CNNs provided a foundation by modeling sequential text and identifying key emotional patterns. However, these methods were limited in capturing long-range dependencies and often required extensive feature engineering.

The introduction of transformer-based models marked a significant breakthrough. Models like BERT (Devlin et al., 2018) introduced bidirectional context representation, enabling a deeper understanding of words by considering both their preceding and following contexts. This innovation proved especially valuable for emotion classification, where contextual nuances are critical. Fine-tuning BERT on emotion-labeled datasets has consistently yielded state-of-the-art results. Following BERT, models such as RoBERTa (Liu et al., 2019) further improved performance by refining training strategies.

More recent advancements, such as DeBERTa (He et al., 2021) and T5 (Raffel et al., 2023), have pushed the boundaries of performance in emotion classification. DeBERTa employs disentangled attention mechanisms to better encode word relationships, enhancing its ability to detect complex emotional signals. T5, with its text-to-text framework, provides flexibility for addressing tasks like multi-label emotion classification, where multiple emotions may coexist within a single text. These transformer models have set a new standard in the field and continue to be the foundation for modern emotion classification research.

1.4 Concept and Work Plan

Phase 1: Data collection and preparation focuses on collecting and preparing raw data from Facebook for analysis. The goal is to obtain messages in JSON format from Messenger exports, clean the data by removing irrelevant metadata and empty messages, and excluding non-text content such as images, links and other multimedia. In addition, a preliminary analysis of the data will be performed to examine the distribution of message lengths and typical patterns, as well as identify potential challenges, such as Polish slang, emoji or spelling errors, that may affect the analysis. The result of this phase is a cleaned message dataset prepared for emotion classification.

Phase 2: In this phase, we will decide how to structure the input data for emotion classification models. Various approaches to segmenting and organizing the text will be explored. For instance, messages could be grouped by conversation threads, time frames, or specific relationships (e.g., all messages exchanged with a particular person, such as a parent or friend). Alternatively, messages could be analyzed as individual texts or aggregated into larger contexts, such as daily or weekly summaries.

By experimenting with different input structures, we aim to evaluate how context affects the models' ability to classify emotions accurately. For example, analyzing a single conversation as a whole might provide richer context for understanding emotional patterns, while segmenting messages by timestamps could highlight temporal changes in sentiment.

Phase 3: Here we will utilize the *Helsinki-NLP/opus-mt-pl-en* model from Hugging Face to translate Polish text into English, ensuring accurate and efficient preprocessing for downstream analysis.

Phase 4: Next we will perform emotion classification on the translated English data using the models *bhadresh-savani\distilbert-base-uncased-emotion* and *SamLowe\roberta-base-go_emotions-onnx*, which classify text into discrete emotions. Additionally, apply the *visegradmedia-emotion\Emotion_RoBERTa_polish6* model to classify discrete emotions in the original Polish text. We will also use two lexicon-based models (one in Polish and one in English) that score single words on dimensions like arousal or valence. Those models also used transformers to extrapolate lexicons to unseen words (Pilisiecki et al., 2023).

Phase 5: The goal of the next phase is to examine the effect of translation on emotion classification and determine which models work better - Polish or English ones. In addition, this phase includes demonstrating the differences in the classification by the two English models - for example, one model identified the message as **sadness** and the other model identified it as **anger**. The result will be a detailed report showing the effects of translation and the differences between the models.

Phase 6: Documentation and presentation focuses on summarizing the project's findings and presenting the results. The final report will include the methodology, results and key findings, supported by visualizations such as emotion distribution charts for Polish and English pipelines, examples of differences in predictions, and the impact of translation on classification quality.

Risk analysis: The risk analysis for the project highlights three key challenges. First, the lack of labeled ground truth makes it difficult to assess prediction accuracy; this is mitigated by comparing the pipeline results of two models and assessing their intuitive consistency with the message content. Second, translation can change the emotional context of a message, which is addressed by manually reviewing and documenting significant changes. Finally, cultural differences in emotional interpretation can lead to discrepancies, mitigated by relying on intuitive assessments of emotional context in Polish to guide analysis.

Note: For obvious and legal reasons, we keep messages private, without revealing names.

1.5 Approach & Research Methodology

The final evaluation of the project will focus on both quantitative and qualitative analyses. Indicators include the distribution of emotions, the consistency of predictions between pipelines, and the effect of translation on performance. Qualitative analysis examines edge cases where predictions differ, and translation-induced changes in emotional context are checked manually. Visualization tools, such as emotion distribution charts (e.g., how often a person feels **sadness** versus how often **joy**), help illustrate the results.

The project uses the Hugging Face Transformers library and PyTorch for model implementation and inference, Pandas and Numpy for handling the data, visualization tools such as Matplotlib. Computing resources include local CPUs and GPUs or cloud platforms such as Google Colab.

References

Paul Ekman 1992. *An argument for basic emotions*. Routledge *Cognition and Emotion*, 169–200

James Russel 1980. *A circumplex model of affect*. Journal of Personality and Social Psychology, 6(39):1161–1178

Pennebaker, James and Francis, Martha and Booth, Roger 1999. *Linguistic inquiry and word count (LIWC)*.

Saif M. Mohammad and Peter D. Turney 2013. *Crowdsourcing a Word-Emotion Association Lexicon*.

Strapparava, Carlo and Valitutti, Alessandro 2004. *WordNet-Affect: an Affective Extension of WordNet*. Vol 4., 4

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. *EmpaTweet: Annotating and Detecting Emotions on Twitter*. European Language Resources Association (ELRA). *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3806—3813

Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. 2014. *EMOTEX: Detecting Emotions in Twitter Messages*.

CBalabantaray R., Mudasir Mohammad, Sharma Nibha. 2016. *Multi-Class Twitter Emotion Classification: A New Approach*. International Journal of Applied Information Systems, 4. 48-53. 10.5120/ijais12-450651

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 10.48550/arXiv.1810.0480

Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 10.48550/arXiv.1907.11692

Pengcheng He and Xiaodong Liu and Jianfeng Gao and Weizhu Chen 2021. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. abs/arXiv.2006.03654

Colin Raffel and Noam Shazeer and Adam Roberts and Katherine Lee and Sharan Narang and Michael Matena and Yanqi Zhou and Wei Li and Peter J. Liu 2023. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. abs/1910.10683

Plisiecki, H., Sobieszek, A. Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. Behav Res 56, 4716–4731 (2024). https://doi.org/10.3758/s13428-023-02212-3

Plisiecki, H., Sobieszek, A. 2024. *Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection.*. Behav Res, 56, 4716-–4731