UNIVERSITY OF
LATVIA

Summary of
Doctoral Thesis

Viktorija Ļeonova

EMPLOYING MACHINE LEARNING TECHNIQUES FOR DETERMINING SPEAKER'S FRUSTRATION LEVEL IN SOCIAL MEDIA TEXTS IN LATVIAN

Riga 2025

UNIVERSITY OF
LATVIA

FACULTY OF SCIENCE AND TECHNOLOGY

Viktorija Ļeonova

EMPLOYING MACHINE LEARNING TECHNIQUES FOR DETERMINING SPEAKER'S FRUSTRATION LEVEL IN SOCIAL MEDIA TEXTS IN LATVIAN

SUMMARY OF THE DOCTORAL THESIS

Submitted for the degree of Doctor of Natural Sciences
in Computer Science
Subfield: Intelligent Systems Theory

Riga 2025

The doctoral thesis study was carried out: at the Department of Computer Science, Faculty of Computer Science, University of Latvia, from 2016 to 2024.

The thesis contains an introduction, 9 body sections, reference list, 11 appendices.

Form of the thesis: dissertation in Computer Science, in Intellectual System Theory subfield.

Supervisor: Prof., *Dr. sc. comp.* **Jānis Zuters**, University of Latvia.

Reviewers:

1. Prof., *Dr. sc. comp.* **Inguna Skadiņa**, University of Latvia;
2. Prof., *Dr.* **Tomas Krilavičius**, Vytautas Magnus University, Lithuania;
3. Asist. Prof., *Dr.* **Anton Leontyev**, Geneva College, USA.

The thesis will be defended at the public session of the Doctoral Committee of Faculty of Science and Technology, University of Latvia, at 15:00 on October 17, 2025.

The thesis is available at the Library of the University of Latvia, Kalpaka blv. 4.

Chairman of the Doctoral Committee
of Faculty of Science and Technology,
University of Latvia _____/ Guntis Bārzdiņš /

Secretary of the Doctoral Committee _____/ Sintija Siliņa /

ISBN 978-9934-36-434-1
ISBN 978-9934-36-435-8 (PDF)

Abstract

This study introduces a novel model for automatic annotation of social network dialogues and prediction of the user's emotional state dynamics. This model utilises an unprecedentedly comprehensive list of non-lexical means of expression along with the traditional bag-of-words input. This study also presents the adaptation of the model to Latvian and demonstrates that the adapted model performs equally well in English and Latvian and is thus language-independent. This study additionally introduces and describes a new dataset in Latvian that was constructed specifically for this purpose. This new dataset consists of 300 dialogues between users and customer support representatives collected from social media, specifically, from the *X*[1] social network, in which the dissatisfaction (frustration) level is annotated at the user message level on a scale of 0 to 4, and the dynamics of emotional state in the dialogue are computed from the change of emotional state after the technical support specialist's response.

By using non-lexical means of expression, the model accurately annotates the level of frustration in messages and predicts changes in emotional state dynamics. The results demonstrate the effectiveness of incorporating non-lexical means of expression in improving the accuracy of emotion detection and prediction. This study has potential applications in improving customer service and enhancing user experience in social media.

[1] At the time of data collection — Twitter

Table of Contents

Abbreviations Used in This Summary

NLME: Non-Lexical Means of Expression
LLM: Large Language Model

1. Introduction

To be able to lead a productive and functional life in a society, one needs to be able to measure their relationship with others. No coordination or collaboration is possible if an individual does not have a theory of mind of another, or in other words, is unable to predict his or her behaviour. Such behaviour, in turn, is strongly dependent on the attitude towards the objects and phenomena of the environment, including other individuals.

This principle holds for any social species, and of course, it holds for humans. Some scientists speculate (Whiten, de Waal, 2017) that our brain has developed because of our extensive social interactions, to navigate an ever-changing social landscape in a closed group. Whether this is true or not, emotions and their recognition in others play a vital part in our lives. It is only natural that with the advent of the Internet, especially Web 2.0 with its abundance of user-generated content, researchers would seek to try and formalize the recognition of emotions in digital media.

The very volume of such media makes it nearly impossible for humans to feasibly work with or even browse through it all, as described in Verma et al. (2016). Thus, researchers must find viable methods for their automatic processing.

Fortunately, the same tremendous increase in processing power, storage volumes, and bandwidth that has allowed the users to generate unparalleled volumes of various media content has provided the means for developing technologies for harnessing this content. And so, researchers have continuously sought to employ the most advanced techniques to annotate emotions in user-generated content. For example, this has enabled research into determining the general sentiment toward a public person or a phenomenon (Wang et al, 2020) or into developing an emotion-aware healthcare system (Ayata et al, 2020).

1.1. Research Area

This thesis describes research in the area of automatic emotion annotation, with a particular focus on frustration detection in English and Latvian social media texts. This is a multi-faceted area connected to machine learning, linguistics and psycholinguistics, neuroscience and many other fields.

While existing research papers in this area focus on models that primarily classify basic emotions, frustration — an emotion frequently experienced in communication — remains largely overlooked. Given that companies rely more and more on social media to interact with customers, accurate detection of frustration could serve as a valuable tool indicating user satisfaction or dissatisfaction. However, research on frustration recognition, especially in textual sources, remains scarce.

Most existing emotion annotation approaches rely on lexical analysis, utilizing word sequences, n-grams, and sentiment lexicons to determine emotional tone. While effective, these methods largely ignore non-lexical features, which are often encountered in informal digital communication. Social media messages, unlike structured written texts, incorporate diverse non-lexical elements, including punctuation patterns (e.g. excessive exclamation marks, ellipses), emojis (denoting mood or sentiment), and ASCII/Unicode symbols (e.g. ¯_(ツ)_/¯), which contribute significantly to emotional expression. Despite their importance, these features remain underexplored in frustration detection models.

In addition, emotion recognition research is predominantly focused on English, leaving low-resource languages like Latvian at a disadvantage.

1.2. Relevance of the Research Problem

The study described in the thesis seeks to address gaps in the existing research of emotion annotation in text. First of all, such an omnipresent emotion as frustration, knowledge of which could be exploited in a variety of ways such as to measure customer satisfaction and dissatisfaction, have not been studied sufficiently. Another important aspect is that such promising text elements as non-lexical means of expression (NLMEs) are underutilised and no comprehensive list of them is available. At the same time, complex n-grams and metrics are widely used for lexical means of expression but primarily for major languages, so this work also covers the aspect of availability of tools for low-resource languages. Specifically, it focuses on Latvian, with an ambition of creating a language-independent tool that would not require an immense corpus of marked texts to be useful.

This thesis seeks to fill in these gaps. The underrepresented emotion of frustration is studied and annotated (Leonova, Zuters, 2021), and in doing so, a new dataset for the low-resource language of Latvian is created, adding to the number of available corpora. This work also evaluates the understudied textual features of NLMEs and demonstrates that the addition of features based on NLMEs would improve the accuracy of frustration recognition. It also shows that this hypothesis holds for both Latvian and English datasets.

1.3. Aim and Objectives

The undertaken research aims to create a model that will accurately predict the level of frustration expressed in a user's message in his or her dialogue in English or Latvian with a customer support representative based on the lexical and non-lexical features of the message, or based on the lexical and non-lexical features in the user's previous message and the customer support representative's

message directly following it. This study aims to achieve these goals through the following tasks:

Find or develop a suitable emotion annotation model that accounts for frustration levels in an excerpt of text, such as a social media message.

1. Construct a neural-network-based model that is able to automatically annotate levels of frustration in messages based on the text of the message.
2. Find or develop a dataset suitable for training the frustration annotation model.
3. Train and test the model.
4. Study the dataset and identify NLMEs that could potentially serve as predictors for the level of frustration expressed in a message.
5. Compile the list of NLME-based features demonstrating correlation with the level of frustration expressed in the message.
6. Construct a model employing the NLME-based features as input parameters, and test their effect on the model's performance.
7. Examine the dataset for potential employment of data processing techniques, and test the effect of these techniques on the resulting model performance.

Successfully solving these tasks will:

1. Facilitate data processing for Latvian messages in social media.
2. Provide a baseline for frustration annotation tasks in English and Latvian.
3. Possibly benefit emotion annotation tasks in other languages sharing the same cultural paradigm and similar use of NLMEs.
4. Provide a comprehensive list of NLMEs found in social media messages in English and Latvian.

1.4. Theses and Research Questions

This study examines the following theses:

1. In customer support dialogues in English and Latvian, the intensity of frustration expressed in a user's message and measured as a median of grades assigned by three independent annotators can be predicted from the text of the user's message using a neural-network-based model with adaptive vocabulary construction. Prediction accuracy is further improved by adding NLME-based metrics, close-to-morphological text segmentation and further text processing such as diacritics removal, and the results are comparable to results achieved by large language models (LLMs).
2. In customer support dialogues in English and Latvian, the frustration intensity of the user's message can be predicted from keywords in the user's previous message together with keywords (from a different set) in the intervening message from customer support.

The following research questions were investigated:

1. How to measure the intensity of frustration expressed in social media messages?
2. What characteristics do datasets require to be usable for model training and testing?
3. What NLME-based metrics can be used for frustration prediction?
4. What is the best frustration prediction model configuration?
5. What data processing techniques can be employed to improve prediction accuracy?
6. How does the prediction accuracy compare to the performance of general-purpose LLMs?

1.5. Research Methods

This study was performed using the following research methods:

- *Review of scientific literature.* Peer-reviewed sources including books, journals, and publications, as well as arXiv preprints in the field of emotion annotation, neural networks, and psycholinguistics were studied in order to identify baseline and cutting-edge approaches in emotion annotation. A knowledge gap was found: there was a lack of research on frustration annotation and on using NLMEs in emotion annotation in Latvian and English.
- *Iterative development process.* Tools and models developed for the purpose of this study were designed, implemented, and refined iteratively. Improvements and changes were implemented following the evaluation of results from preceding iterations.
- *Controlled experimental studies.* The performance of algorithms and neural-network models was systematically assessed through controlled experimental setups.
- *Quantitative evaluation.* The evaluation techniques used in this thesis were based on numerical metrics, enabling objective comparison of different approaches and models.
- *Error analysis.* Where possible, errors were analysed in order to identify potential causes and possible improvements.

1.6. Main Results

This thesis presents the following developments in frustration recognition in social media messages:

- Interactive vocabulary construction: a novel iterative approach to building a predictor dictionary was implemented, allowing model training on small annotated datasets.

- New approach to emotion annotation: unlike traditional models relying solely on lexical features, this work integrates NLME-based features, such as punctuation patterns, emojis, and other features characteristic of online communication, to improve frustration level annotation.
- New datasets for Latvian and English[2]: two new datasets consisting of user dialogues with customer support representatives annotated for frustration levels (one for English and one for Latvian) were developed, increasing the number of publicly available corpora. The English dataset was derived from Kaggle's Twitter Customer Support (Axelbrooke, 2017) dialogues. The Latvian dataset was built manually from scratch.
- Text preprocessing: this study demonstrates that close-to-morphological segmentation slightly improved accuracy (+1pp) by reducing word-form variability.

The proposed NLME-enhanced model significantly outperforms baseline models:

- Latvian dataset: +18pp improvement over baseline, +7pp over lexical-only model.
- English dataset: +19pp improvement over baseline, +6pp over lexical-only model.
- The NLME-only model (language-independent) achieved 40% accuracy, performing comparably to the lexical-only model (41%), demonstrating the effectiveness of non-lexical features.

The presented models vs. Large Language Models (LLMs):

- GPT-3.5 Turbo and llama 1:8B performed at or below baseline.
- GPT-4o-mini reached lexical-only model accuracy but did not surpass the proposed model.
- gemma 2:9B outperformed the proposed model in English (+7pp) but lagged in Latvian (−4pp).
- GPT-4o surpassed the proposed model (+3pp in Latvian, +8pp in English), but as a paid large-scale model, it lacks the adaptability of the proposed lightweight, locally trainable model.

Additional results:

- Customer support responses had little impact on reducing frustration levels, showing only a 0.34-point decrease on average between turns.
- Automatic removal of diacritical marks and spelling normalisation had no meaningful effect on accuracy.
- NLME features proved to be highly predictive, with accuracy comparable to traditional lexical approaches.

This thesis shows that integrating NLMEs significantly enhances frustration recognition and offers a scalable, adaptable alternative to LLMs, particularly for low-resource languages like Latvian. It also demonstrates that NLME features derived from the Latvian dataset, when added to the frustration

[2] Datasets available at https://github.com/Lynxa/AnnotatedTweets

recognition model, are similarly beneficial for the English dataset. Thus, the model is effectively language-independent (but not culture-independent) and can be employed for frustration recognition in English and potentially in other languages sharing the same cultural paradigm. For yet other languages, the model may require studying and deriving the NLMEs used by the bearers of respective cultures.

1.7. Approbation and Publication of the Author's Works

The author contributed to the following peer-reviewed publications pertaining to this thesis:

- Leonova, V., 2020. Review of non-English corpora annotated for emotion classification in text. In *Proceedings for Databases and Information Systems: 14th International Baltic Conference, (DB&IS 2020), 14* (pp. 96-108). Springer International Publishing.
- Zuters, J. and Leonova, V., 2020. Adaptive Vocabulary Construction for Frustration Intensity Modelling in Customer Support Dialog Texts. *International Journal of Computer Science & Information Technology (IJCSIT) 12*.
- Zuters, J. and Leonova, V., 2020, November. Frustration Intensity Prediction in Customer Support Dialog Texts. In *Proceedings for CS & IT Conference 10(14)*. CS & IT Conference Proceedings.
- Zuters, J., Strazds, G. and Ļeonova, V., 2019. Morphology-Inspired Word Segmentation for Neural Machine Translation. *Databases and Information Systems X* (pp.225-239).
- Leonova, V. and Zuters, J., 2021, September. Frustration Level Annotation in Latvian Tweets with Non-Lexical Means of Expression. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 814-823).
- Leonova, V. and Zuters, J., 2022. Frustration Level Analysis in Customer Support Tweets for Different Languages. In *Proceedings for 15th International Baltic Conference on Digital Business and Intelligent Systems (DB&IS 2022) Forum*.
- Leonova, V. and Zuters, J., 2022. Frustration Level in Customer Support Tweets: Towards a Language-Independent Model. *Baltic Journal of Modern Computing, 10(4)* (pp.738-753).

Four publications that served as a foundation for the thesis, were presented in the following scientific conferences:

- Leonova, V. Review of non-English corpora annotated for emotion classification in text. Databases and Information Systems: *14th International Baltic Conference, DB&IS 2020, Tallinn, Estonia, June 16–19, 2020.*

- Zuters, J. and Leonova, V., Frustration Intensity Prediction in Customer Support Dialog Texts. 9th International Conference on Information Technology Convergence and Services ITCS 2020, Sydney, Australia, December 19-20, 2020.
- Leonova, V. and Zuters, J., 2021. Frustration Level Annotation in Latvian Tweets with Non-Lexical Means of Expression. *International Conference on Recent Advances in Natural Language Processing (RANLP 2021), online, September 1-6, 2021.*
- Leonova, V. and Zuters, J., Frustration Level Analysis in Customer Support Tweets for Different Languages, 15th International Baltic Conference on Digital Business and Intelligent Systems DB&IS, Riga, Latvia, July 03-06, 2022.

This research was approbated in the following research project:
- Joint project of SIA TILDE and the University of Latvia "Multilingual Artificial Intelligence Based Human Computer Interaction" No.1.1.1.1/18/A/148 supported by the European Regional Development Fund.

2. Summarising Description of the Promotion Work

2.1. Background

It is universally known that the Internet has emotionally charged content, harnessing which could greatly help in the development of recommendation systems, chatbots, and other tools. However, automated emotion recognition in text presents challenges due to its subjective nature and the limitations of capturing emotional nuances through words alone. Punctuation, such as the absence of periods in internet jargon, adds complexity (Khalifa, 2020).

While emotion recognition initially focused on speech, the rise of social networks like Facebook and X shifted the emphasis to text (Dean, 2023). However, a significant observation is the prevalent reliance on Ekman's base emotion system (Ekman, 1992) in non-English datasets (Leonova, 2020). This system, proposing six basic emotions, is criticized for its simplicity and arbitrariness (e.g. Gendron et al., 2014), raising questions about its applicability.

As machine learning, particularly neural networks, advanced, emotion recognition extended to text (Moriyama, Ozawa, 1999). Despite this progress, a notable gap exists in the aspect of frustration — a universally recognized emotion highly relevant to customer support and service quality assessment (Stauss et al., 2005). Existing studies on frustration annotation are limited and dated (Klein et al., 2002; Hone, 2006), with recent works focusing on different goals and methods (Hu et al., 2018).

Emotion classification typically relies on words or lexical means, leading to lexicons. While some explore non-lexical features like emojis or hashtags, their limited presence in certain contexts hinders applicability. Non-lexical features are more advanced in voice communication, lacking a systematic review in text-based emotion recognition.

Most state-of-the-art models concentrate on English, with few available to low-resource languages. This highlights a research gap in extending emotion recognition models to diverse linguistic contexts (Gruzitis et al., 2018).

This promotional work takes the best of both worlds and addresses the problem of recognising frustration with the help of a proposed neural-network model. The design of the study is described next.

2.2. Datasets

For experiments, two datasets were used. First was an English dataset of 376 dialogues. Each user's turn was rated by three annotators, and only "valid" turns — those receiving numeric values from all three experts — were used in experimentation. The final dataset consisted of 843 annotated valid user turns and 470 valid support turns. The second was a Latvian dataset, consisting of 283 dialogues totalling 688 user turns and 398 customer support representative turns.

The building of both datasets is described in the section “Data Collection and Preparation”.

2.3. Data Collection and Preparation

English. The dataset was built using the Kaggle Customer Support Twitter dataset, by linking replies to message IDs and creating consecutive dialogues containing interactions between users and customer support representatives. A subset of 400 dialogues was selected for annotation, ensuring at least 800 customer turns. Dialogues were anonymised, and user/support IDs were replaced with labels. Human annotators assigned symbols representing frustration levels to customer turns, with possible values ranging from 0 to 4, “n” denoting uncertainty, and empty values indicating incomprehensible text.

To evaluate the consistency of the annotators’ ratings, intraclass correlation was measured, indicating moderate to good consistency for both languages.

Latvian. The dataset contains conversations between users and customer support representatives from four major Latvian Internet and telecommunication service providers: Tet, LMT, Bite, and Tele2. The data was collected from X accounts associated with these providers. Each conversation includes at least two user turns with a customer support turn in between, and the dialogues are in Latvian. The dataset, manually collected to ensure adherence to criteria, consists of 283 dialogues, providing insights into user–customer support interactions in the X social network. In the same manner as in the English dataset, dialogues were anonymised and user names and e-mails were substituted by respective labels. Three annotators assigned perceived frustration levels to the user messages according to the same grading scheme as in the English dataset.

2.4. Data Processing

During this step, the effect of preprocessing user messages in Latvian and English on the accuracy of frustration prediction with the proposed model was examined. Irregularities commonly encountered in colloquial language were identified, and possibilities were explored for their unification, based on the idea that it might improve model accuracy. The whole preprocessing pipeline includes segmentation, removal of diacritical marks, abbreviation unification, and normalisation of colloquial language.

Segmentation was applied to Latvian messages due to the language’s inflectional nature, hypothesizing that breaking words into close-to-morphological components could reduce the number of grammatical forms and accordingly enhance model performance. The removal of diacritical marks was also explored, addressing common issues in social media text where users omit or misuse diacritical marks due to technical limitations or lack of grammatical awareness. An attempt was made to replace the most frequently used

colloquialisms and to unify abbreviations, but only segmentation had a visible effect on the performance of the proposed model.

NLMEs in the datasets, such as punctuation (exclamation marks, question marks, and ellipses), words written entirely in the upper case, repeated letters, emojis, hashtags, and message length, were identified and analysed for their correlation with frustration intensity. While message length proved to be the strongest predictor, features such as punctuation and references to authorities (e.g. mentions of the Consumer Rights Protection Centre) also showed positive correlations. However, some elements, like hashtags, did not contribute to frustration prediction.

In conclusion, using the combination of NLME-based input parameters and data segmentation allows significantly improving the accuracy of predictions.

2.5. Study

2.5.1. Experimental Setup

For the experiments, two corpora — English and Latvian — of X social network dialogues between a user and a support representative were used. Each user's turn was rated by three annotators, and for experimentation purposes, only the "valid" turns —ones to which each of the experts had assigned a numerical value — were utilised. Some turns were difficult to evaluate, so the final version of the data consisted of 843 (688 for Latvian) annotated valid user turns, as well as 470 (398 for Latvian) valid support turns, that is, messages located between two valid user turns. As for the frustration dynamics in the dialogues, the average frustration intensity change from one user turn to the next was −0.35, which means that, in general, frustration intensity was observed to decrease from turn to turn, but only slightly. Over the course of a short dialogue, the user's frustration intensity rating, on average, was expected to remain essentially unchanged.

For the turn's text encoding, two keyword vocabularies were constructed in the form of lists of lower-cased keywords that occurred in the valid turns at least three times, separately for user turns and for support turns. The keywords were ordered by the certainty of their rating, and only a number of the best keywords was kept for the construction of bags of words:

- for user turns, the keywords were ordered by the standard deviation of ratings of the turns the particular keyword occurred in;
- for support turns, the keywords were ordered by the standard deviation of the difference of ratings of the two user turns surrounding every support turn that the particular keyword occurred in.

Thus, two vocabularies — the users keyword vocabulary and the support keyword vocabulary — were obtained. Individual messages were represented in

the bag-of-words encoding using the best keywords from the vocabulary, with a 1 or 0 for each keyword denoting its presence or absence in the message:

- encoding was done separately for user turns and support turns,
- a different number of "best keywords" was used in different experiments: the number of keywords was defined by the actual experiment configuration.

Additionally, the datasets were examined and all the NLMEs identified. Those that showed some correlation between the NLME-based metric and the frustration score were used as input parameters.

The prediction model was built as a linear neural network taking the bag of words and (in the advanced model) the NLME-based metrics calculated from the corresponding social media message and producing a predicted frustration level on the discrete scale of 0 to 4 as its output. The actual model parameters and their selection were adjusted in the course of experiments described below.

The prediction model's performance was evaluated using accuracy metrics. As the prediction classes [0..4] are ordered, a weaker evaluation metric was also introduced: accuracy with tolerance ±1 (such that an "off-by-one" prediction is also considered correct). Accuracy with tolerance seems fitting, seeing as e.g. predicting 2 when the "true" rating is annotated as 3 and predicting 0 in the same situation are not equally wrong.

2.5.2. Phase 1: Dynamics in Dialogue (English Dataset)

- Objectives: This phase aimed to examine hypotheses concerning the predictability of user frustration intensity in customer support dialogues. Two main hypotheses were tested, involving the prediction of frustration from user messages and the impact of customer support responses on user frustration dynamics.
- Methodology: A corpus of dialogues, annotated for frustration levels, was utilized. The study employed keyword vocabularies, bag-of-words encoding, and prediction models to analyse the dynamics of frustration in customer support interactions.
- Results: The study observed a slight decrease in frustration intensity from turn to turn, and awareness of customer support responses did not help in predicting the user's next emotional state.

This phase of the study focused on examining two hypotheses related to dynamics in customer support dialogues using an English dataset. The hypotheses investigated the predictability of user frustration intensity based on the text of user messages and the impact of customer support responses on the user's emotional state. The results demonstrated the proposed neural network–based model's ability to predict user frustration intensity from the text of user messages. The model outperformed the baseline, which assigned the most frequent label to all messages, achieving 71% accuracy with tolerance ±1.

The study also explored frustration intensity dynamics by predicting changes in the user's emotional state based on the support agent's reply.

However, the results indicated that knowing the contents of the support agent's message did not significantly contribute to predicting changes in the user's frustration state.

2.5.3. Phase 2: Frustration Level and Frustration Dynamics in Dialogue (Latvian Dataset)

- Objectives: Extending phase 1 to the Latvian dataset, this phase tested similar hypotheses in the context of Latvian customer support dialogues.
- Methodology: A new dataset in Latvian was constructed and processed similarly to the English dataset. The study compared the results to the English dataset to assess the transferability of findings between languages.
- Results: The study found similarities in the performance of frustration prediction models in the English and Latvian datasets, indicating that the proposed approach could be effective across languages.

In this phase, the study scope extended to frustration levels and dynamics in Latvian customer support dialogues. Hypotheses mirroring those from phase 1 were tested on a newly constructed Latvian dataset. The adaptation of the model involved exploring various configurations of meta-parameters and data processing techniques.

The Latvian dataset, collected from major service providers' X accounts, underwent similar processing to the English dataset. The proposed model's performance was assessed using accuracy measures, with the baseline set at the most frequent value.

Results indicated comparable accuracy for the unmodified model on the Latvian dataset, with a slightly lower gain relative to the baseline. Detailed findings and comparisons are available in the study's relevant sections.

2.5.4. Phase 3: Model Extension with Data Processing and NLME Features (Latvian)

- Objectives: This phase explored the impact of word segmentation, unification of messages, and the inclusion of NLME features on the accuracy of frustration prediction models in Latvian.
- Methodology: Different configurations involving NLME features and various data processing techniques were tested. The study aimed to optimize the model's performance by incorporating NLME information and processing textual data.
- Results: The inclusion of NLME features such as message length and punctuation usage improved the accuracy of frustration prediction models. Data processing techniques such as segmentation further enhanced model performance.

In this phase, the study focused on extending the model with data processing and NLME features for frustration intensity prediction in Latvian customer support dialogues. Hypotheses related to the effect of grammar, diacritical marks,

acronyms, and NLMEs on the resulting model accuracy were analysed. The impact of NLME features and various data processing techniques on model accuracy was explored.

Results showed that segmentation slightly improved accuracy, while abbreviation replacement had mixed effects. NLME features, including message length, exclamation marks, question marks, and periods, contributed positively to accuracy. The final model included all identified NLME features.

The study also examined NLME features for frustration intensity prediction in English messages, finding similarities between Latvian and English in terms of feature impact. The best-performing features were identified, and the model achieved comparable results in both languages. Additionally, meta-parameters like hidden neurons and vocabulary size were explored. Segmented input with 64 hidden neurons and 100-keyword vocabularies yielded the highest performance, demonstrating a 7pp improvement over the old model.

In this phase, the model was successfully extended by incorporating NLME features and optimizing meta-parameters for improved accuracy on both the Latvian and the English dataset.

2.5.5. Phase 4: Language-Independent Model

- Objectives: This phase investigated the development of a language-independent model where lexical information was excluded from the model input parameters and the model was run with a zero-size dictionary.
- Methodology: The study tested the language-independent model with different numbers of hidden neurons to assess its effectiveness in predicting frustration intensity across languages without utilizing language-specific lexical material.
- Results: The language-independent model showed promise as it relied solely on NLME metrics for predicting frustration, providing a potential approach for cross-language applicability.

In this phase, the study explored the development of a language-independent model by running it with a dictionary size equal to zero. This modification transformed the model into a new version that relies solely on NLME-based metrics as input parameters, eliminating the dependence on lexical material from messages for determining frustration. The model was tested with varying numbers of hidden neurons, and the best performer was identified. The study assessed the performance of this language-independent model, exploring its effectiveness in predicting frustration intensity without relying on language-specific lexical information.

These phases collectively form a comprehensive exploration of predicting and understanding frustration dynamics in customer support dialogues, considering both language-specific and language-independent aspects.

2.6. Comparison with LLMs Used for Message Annotation

For evaluation of the effectiveness of the proposed model, it was compared against five prominent LLMs, namely, GPT-3.5, GPT-4o mini, GPT-4o, llama, and gemma. In the experiments, these models were used in the following versions: GPT-3.5 Turbo (OpenAI, 2022), GPT-4o mini (OpenAI, 2024), GPT-4o (OpenAI, 2024), llama3:8b (Ollama, 2024), llama3:8b-uncensored (Sun, 2024), gemma1:7b (Ollama, 2024), gemma2:9b (Ollama, 2024).

For results to be comparable, the same datasets were used as a source of messages and their respective grades assigned by annotators. The LLMs listed above were in turn used to annotate the messages for frustration levels using prompts specifically constructed and tailored for this purpose, and the results were compared to the performance of the proposed model.

The following table (Table 1) demonstrates the performance of gemma2:9b, GPT-3.5 Turbo, GPT-4o mini, and GPT-4o in comparison with the proposed NLME-enabled model. llama 3:8b and llama 3:8b-uncensored run on the Latvian dataset demonstrate a performance of 27% and 22%, respectively, while gemma1:7b correctly predicts the grades in 13% of cases, not even reaching the baseline, thus these LLMs were excluded from further experimentation.

It can be seen that GPT-3.5 Turbo is not up to the task, being almost precisely at the level of the baseline. GPT-4o mini performs better, achieving the level of the proposed model without NLMEs for English data and almost achieving the performance of the improved model for Latvian (1pp difference). GPT-4o and gemma2:9b achieve better results, showing that LLMs can detect frustration better than our proposed model.

Table 1. Performance of LLMs trained and tested on English and Latvian data

Model	Accuracy in comparison with human annotators for English, %	Accuracy in comparison with human annotators for Latvian, %
gemma2:9b	56	44
gpt-3.5-turbo	28	33
gpt-4o-mini	39	48
gpt-4o	57	52
baseline	28	31
model w/o NLME	42	41
proposed model with NLME	49	49

However, the results achieved are comparable, with GPT-4o demonstrating a lead of 3pp for the Latvian dataset and 8pp for English, and gemma2:9b having a similar lead of 7pp for English, but only performing slightly better (3pp) than the model without NLMEs in Latvian. A possible explanation is that the materials used for its training were in English, and not in Latvian. At the same time, it can be seen that the results demonstrated by the proposed model can be replicated by a general-purpose LLM such as GPT-4o, indirectly validating the achieved results.

2.7. Results

The main results presented in this work include:

1. Two new datasets, in Latvian and English, respectively.
2. A comprehensive list of NLMEs employed by social media users in their conversations with customer support representatives.
3. A neural network–based model for annotating user messages with perceived levels of user frustration.

2.7.1. Datasets

This study introduces and describes two datasets in Latvian and English, containing customer support dialogues used to train and evaluate a neural network model for frustration prediction. The English dataset, derived from the Kaggle Twitter Customer Support dialogues, includes nearly 400 conversations and 900 messages annotated for frustration intensity (0-4). The Latvian dataset, manually collected from major service providers on X, consists of 283 dialogues with 688 user turns and 398 support responses, similarly annotated for frustration levels.

2.7.2. List of NLMEs

This study presents a list of NLMEs that were identified upon studying the datasets. The employment of those NLMEs can increase the accuracy of emotion annotation, as we demonstrate for perceived levels of frustration. The NLMEs are listed in Table 2.

Table 2. Identified NLMEs and their correlation with the annotated level of frustration

Feature	**Correlation with frustration level median value**
Number of exclamation marks in the message	a very weak positive correlation (0.05)
Number of exclamation points in the message normalized by message length	a weak negative correlation (-0.10)
Number of question marks in the message	a weak positive correlation (0.14)
Number or commas in the message	a weak positive correlation (0.26)
Number of dots in the message	a weak positive correlation (0.28)
Number of quotation marks (“, ”, ‘, ’, ") in the message	a weak positive correlation (0.13)

Table 2 — continued

Number of words in the upper case longer than 4 symbols in the message	a weak positive correlation (0.12)
Repeating the letter “a” more than twice in a row in the message	a very weak positive correlation (0.07)
Number of emojis in the message independent of the mood they represent	a weak positive correlation (0.14)
Number of positive smileys (made of typographic marks) in the message	a very weak negative correlation (-0.09)
Number of negative smileys (made of typographic marks) in the message	an extremely weak positive correlation (0.04)
Presence of a picture in the message	a very weak positive correlation (0.05)
Referencing PTACGovLV in the message	a weak positive correlation (0.10)
Message length	an average positive correlation (0.44)

2.7.3. Model

This work presents a neural network–based model with interactive vocabulary construction and the use of NLME metrics as input parameters, which can function language-independently and can annotate frustration levels in a user's message in social media with accuracy significantly above the baseline: for precise annotation, the accuracy is 48% and 49% for English and Latvian respectively, and for approximate annotation, it is 90% for both languages. The high-level model schema is shown in Figure 1.

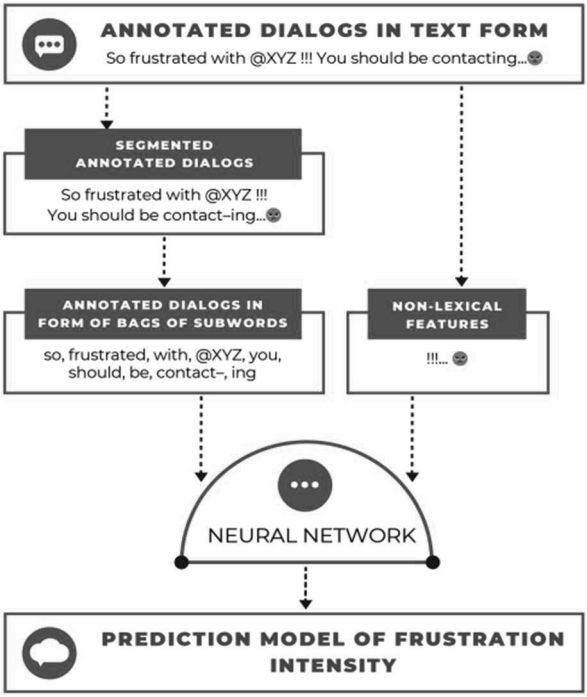

Fig. 1. High-level schema of the frustration level prediction model including NLME-based features as an input

2.8. Discussion and Interpretation

The proposed neural network–based model uses adaptive vocabulary construction and predicts user frustration intensity from the text of a user message addressed at a customer support representative. This model takes an encoded representation of the user's message as input and produces output in the form of an integer rating of frustration intensity on a 5-point scale (0 to 4). The resulting accuracy is compared to a baseline that simply assigns the most frequent label to all instances (28%). In addition to exact accuracy, an accuracy with tolerance 1 (allowing a difference of 1 between the actual and predicted rating) was also calculated. Using this "±1 accuracy" metric, the performance of the model is compared to the appropriate baseline (71% using this metric). Employing adaptive vocabulary construction enables training the model on a small dataset, which is especially important for low-resource languages. For both the English and Latvian datasets, the best results for exact accuracy are achieved using a dictionary containing 100 best predictors and the best results for accuracy with tolerance 1 are achieved using a dictionary of 300 best entries. The best predictors are defined as the words (or word fragments) that demonstrate the lowest standard deviation of annotated frustration level across all messages containing the corresponding word. This training method improves the accuracy of prediction by 13pp and 11pp for exact accuracy for English and Latvian data, respectively, and 9pp for accuracy with tolerance 1 for both languages, which can be ascribed to the baseline being fairly high from the beginning.

This work also describes a series of experiments studying the dynamics of the user's frustration level over the course of the conversation and its correlation with the contents of the customer support representative's messages. The study has concluded that the tested methods did not allow to predict the level of frustration using this approach better than the baseline.

Alongside the interactive vocabulary construction, the model is built using NLME-based metrics as input. The experiments show that these metrics improve frustration level prediction accuracy by another 7pp. Due to the nature of NLMEs, they are only to a limited extent dependent on the language. A comparison of the model's performance on Latvian and English data shows that in both cases the improvement in prediction accuracy is the same. This means that as long as the NLME usage patterns in social media are the same, the model is effectively language-independent.

Data segmentation applied to the datasets also resulted in a similar improvement on Latvian and English data, approximately 1pp. While this was expected from a synthetic language such as Latvian, the improvement demonstrated on the English dataset means that English has a sufficient number of grammatical forms for segmentation to make a difference.

All in all, the results show that using interactive vocabulary construction, NLMEs and segmentation allow to predict the perceived level of frustration on small datasets independently of a specific language.

Comparison with LLMs. The performance of the model was compared to such LLMs as GPT3.5-turbo, GPT 4o-mini and GPT-4o, gemma 1:7, gemma 2:9b, as well as llama 3:8b. The results indicate that GPT-3.5 Turbo, llama 3:8b and gemma 1:7b and GPT4o-mini perform worse than the proposed model, gemma-2:9b gives better accuracy for English and worse for Latvian, and GPT-4o performs considerably (8pp) and slightly (3pp) better for English and Latvian, respectively.

3. Research Applicability Limits

The proposed model is effectively language-independent (but not culture-independent) and can be employed for frustration recognition in English and, potentially, in other languages sharing the same cultural paradigm and therefore using NLMEs in a similar manner. The only prerequisite for this would be the availability of a dataset annotated in a compatible way. As the demonstrated results were achieved on a small dataset, the effort needed for the provision of such a dataset for another language also wouldn't be too taxing.

Naturally, the application of the model to other languages is subject to testing and is currently limited to the means of expression shared between users of European languages. Its extension to other languages is subject to studying and deriving the NLMEs used by the bearers of the respective cultures.

This work is also limited in scope by the studied emotion and medium. The model is only measuring the level of frustration as expressed in texts in social media, even though it stands to reason that other emotions might also correlate with various NLMEs present in the text. In a similar manner, this work only studies a specific type of speech — so-called spoken written text, in particular, Internet speech, found somewhere between oral speech and written text, thus it takes into consideration a specific set of NLMEs extracted from the presented dataset.

4. Main Conclusions and Proposals

This work proposes several developments in the field of frustration recognition and assessment in social media messages.

It presents a new approach towards emotion annotation where, in addition to the emotional charge of the words themselves, NLMEs are employed to determine the emotional state of the message author. The NLMEs used in this work include both conventional means such as exclamation marks, substandard means such as the number of periods in the message, and social media–specific means such as emojis. This study presents a list of NLMEs and explains their

role in denoting frustration. It also demonstrates a novel approach towards using the lexical information of messages, namely, the interactive construction of the best predictor vocabulary. In addition, it proposes using data segmentation to improve the results.

This study proposes a model that uses both NLME-based features, constructed based on non-lexical information found in messages, and a bag-of-words representation, which is derived from the message using an interactively constructed vocabulary, and also employs close-to-morphological segmentation for input processing to make vocabulary construction more coherent. Other methods of input processing, namely, diacritics removal and unification of the most popular variations in spelling, did not yield satisfactory results, even when used simultaneously.

Performance assessment used leave-one-out cross-validation, calculating the accuracy as a percentage of correct predictions obtained by training the model on all data except one entry and predicting the frustration intensity for the one entry left out, repeated for all the entries, and averaged across multiple runs. The proposed model was compared to a model that only had lexical features–based input.

While the model using only the interactively constructed dictionary has demonstrated an improvement of 11pp compared to the baseline, the proposed model applied to the Latvian dataset has achieved an accuracy improvement of 17pp over the baseline and 5pp over the model without NLMEs. Examination of the effect brought in by additional input processing by close-to-morphological segmentation and adjustment of meta-parameters has shown that close-to-morphological segmentation results in an accuracy improvement of 1pp, yielding a total of 18pp improvement over the baseline and 7pp improvement over the lexical features–only model. The same model, when applied to the English dataset, achieves an accuracy of 47%, which gives 19pp and 6pp of increase in accuracy, respectively. However, adjusting the metrics to be less Latvian-specific resulted in achieving accuracy 49% using the English dataset, which is 21pp higher than the baseline and 8pp more accurate than the model without NLME features.

The proposed model was additionally compared to LLMs with the conclusion that it outperforms less-developed LLMs such as llama3:8b and GPT-4o mini but is considerably outperformed on English data by GPT-4o (8pp) and gemma2:9b (7pp), although only slightly outperformed by GPT-4o on Latvian (3p), showing the validity of the achieved results.

As an intermediary development, this study presents a new neural network–based model using interactive vocabulary for frustration intensity prediction for customer messages in their conversations with customer support representatives. Constant assignment of the most popular grade was used as a baseline, and it was demonstrated that the frustration intensity of an individual message can be predicted based on its contents. The accuracy of this prediction is significantly higher than the baseline (41% vs 27% in the case of English and 31% vs 42% in

the case of Latvian). Studying the effect of customer support responses on the emotional state of the user has shown that typically, the user's emotional state mostly remains unchanged, with a small decrease of 0.34 points on average from one turn to the next. Given the challenges in the precise calibration of the user's frustration level, trying to model this dynamic as a function of the emotional valence of the support agent's messages was not met with success in the course of this experiment.

Additionally, another model was tested in the course of the experiments. It only uses NLME-based metrics as input parameters, and it was found that the accuracy of this model (40%) does not differ significantly from the accuracy of the model using interactively constructed vocabulary (41%). It is fully language-independent, and it demonstrates that using lexical and non-lexical content of the message result in similar prediction accuracy.

This study also presents and describes two datasets — in Latvian and English. They represent collections of dialogues in respective languages between users and customer support representatives that contain more than one user message. These datasets were used for training and assessing the described neural network–based model for frustration prediction and for discovering the role and differences of NLME usage in annotating frustration in both languages. The English dataset was developed using a subset of the Kaggle Twitter Customer Support dialogues, consisting of close to 400 dialogues and almost 900 individual messages. The Latvian dataset was collected manually from X accounts of major internet and telecommunication service providers and in total has 283 dialogues with 688 user turns and 398 customer support representative turns. All user messages in both datasets are annotated for frustration intensity on a scale of 0 to 4.

References

Axelbrooke, S., 2017. Customer Support on Twitter [Data set]. Kaggle. Retrieved 9 August 2020 from: https://www.kaggle.com/thoughtvector/customer-support-on-twitter

Ayata, D., Yaslan, Y. and Kamasak, M.E., 2020. Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. *Journal of Medical and Biological Engineering 40*(2) (pp.149-157).

Dean, B., 2023, 27 March. *Social Network Usage & Growth Statistics: How Many People Use Social Media in 2023?* Backlinko. Retrieved 1 November 2023 from: https://backlinko.com/social-media-users

Ekman, P., 1992. An Argument for Basic Emotions. *Cognition & Emotion 6*(3-4) (pp. 169–200).

Gendron, M., Roberson, D., van der Vyver, J.M. and Barrett, L.F., 2014. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion, 14*(2) (p. 251).

Gruzitis, N., Nespore-Berzkalne, G. and Saulite, B., 2018, May. Creation of Latvian FrameNet based on universal dependencies. In *Proceedings of the International FrameNet Workshop (IFNW)* (pp. 23-27).

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J. and Akkiraju, R., 2018. Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).

Klein, J., Moon, Y. and Picard, R.W., 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with computers 14*(2) (pp. 119-140).

Khalifa, N.S., 2019. Internet chat as 'jargon'. *Journal of Language Studies 3*(1) (pp.121-132).

Leonova, V., 2020. Review of non-English corpora annotated for emotion classification in text. In *Proceedings of the Databases and Information Systems: 14th International Baltic Conference, (DB&IS 2020)* (pp. 96-108). Springer International Publishing.

Leonova, V. and Zuters, J., 2021, September. Frustration Level Annotation in Latvian Tweets with Non-Lexical Means of Expression. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 814-823).

Moriyama, T. and Ozawa, S., 1999, June. Emotion recognition and synthesis system on speech. In *Proceedings IEEE International Conference on Multimedia Computing and Systems 1* (pp. 840-844). IEEE.

Ollama, 2024. *Llama 3* [LLM]. Ollama. Retrieved 24 October 2024 from: https://ollama.com/library/llama3:8b

Ollama, 2024. *Gemma* [LLM]. Ollama. Retrieved 24 October 2024 from: https://ollama.com/library/gemma:7b

Ollama, 2024. *Google Gemma 2* [LLM]. Ollama. Retrieved 24 October 2024 from: https://ollama.com/library/gemma2:9b

OpenAI, 2022. *GPT-3.5 Turbo* [LLM]. OpenAI. Retrieved 24 October 2024 from: https://platform.openai.com/docs/models/gpt-3-5-turbo

OpenAI, 2024. *GPT-4o mini* [LLM]. OpenAI. Retrieved 24 October 2024 from: https://platform.openai.com/docs/models/gpt-4o-mini

OpenAI, 2024. *GPT-4o* [LLM]. OpenAI. Retrieved 24 October 2024 from: https://platform.openai.com/docs/models/gpt-4o

Stauss, B., Schmidt, M. and Schoeler, A., 2005. Customer frustration in loyalty programs. *International Journal of Service Industry Management*, *16*(3) (pp.229-252).

Sun, J. [sunapi386], 2024. *Ollama import of Orenguteng/Llama-3-8B-Lexi-Uncensored-GGUF* [LLM]. Ollama. Retrieved 24 October 2024 from: https://ollama.com/sunapi386/llama-3-lexi-uncensored

Verma, J.P., Agrawal, S., Patel, B. and Patel, A., 2016. Big data analytics: Challenges and applications for text, audio, video, and social media data. *International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI) 5(1)* (pp. 41-51).

Wang, S., Schraagen, M., Sang, E.T.K. and Dastani, M., 2020, December. Public sentiment on governmental COVID-19 measures in Dutch social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Whiten, A. and van de Waal, E., 2017. Social learning, culture and the 'socio-cultural brain' of human and non-human primates. *Neuroscience & Biobehavioral Reviews, 82* (pp. 58-75).

Acknowledgements

The research used in this paper has been in part supported by the European Regional Development Fund within the joint project of SIA TILDE and the University of Latvia "Multilingual Artificial Intelligence Based Human-Computer Interaction" No.1.1.1.1/18/A/148.

The author would like to express her gratitude: to Dmitry Kuzmenko, M.Sc. in CS, for code review, to Ekaterina Alekseyenok, M.H. in Philology, Maria Bogina, Cognitive Studies B.A. for consulting in linguistics and psycholinguistics; to Svetlana Burmistrova, Sp. in Graphic Design for helping with graphical schemas; to Katerina Litvinova, B.Soc.Sc. in Psychology, to Jelena Burtseva, B. Econ.Sc., to Svetlana Shchegolihina, for help in the annotation of the data set to Oleg Oshmyan, M.Eng. in CS, for editing, to Aleksandrs Rivoss, M.Sc in CS, for everything.