Baltic J. Modern Computing, Vol. 10 (2022), No. 4, 738-753
https://doi.org/10.22364/bjmc.2022.10.4.08

Frustration Level in Customer Support Tweets: Towards a Language-Independent Model

Viktorija LEONOVA, Jānis ZUTERS

University of Latvia, Raiņa bulvāris 19, Riga, LV-1586, Latvia

{viktorija.leonova, janis.zuters}@lu.lv

Abstract. In this paper, we present a comparative analysis of frustration intensity prediction for tweets in different languages using neural-network-driven models combining lexical and non-lexical means of expression. The different configurations of models were tested on customer support dialog texts in two languages – Latvian and English. We show that our model is effectively language-independent within the same culture. The experimental results show the texts in both languages to be effectively evaluated for frustration intensity with slightly better overall results in Latvian. For both languages, the prediction models with configurations using all available features based on non-lexical means of expression yield the best accuracy, while the utilization of those features result in similar improvement in both languages.

Keywords: Machine learning, deep learning, neural networks, emotion annotation, frustration, non-lexical means of expression

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 model of another, in other words, is unable to predict their behaviour, which, accordingly, is strongly dependent on the attitude towards the objects and phenomena of the environment, including other individuals. For humans, as a social species, this is also true. Some scientists speculate (Whiten and de Waal, 2017) that our brain has developed because of our extensive social interactions for the purposes of navigating an ever-changing landscape in a closed group. Whether it is true or not, emotions and their recognition in others play a vital part in our life. And it is only natural that with the high noon of the Internet, especially Web 2.0 with its abundance of user-generated content, the researchers would seek to try and formalize the recognition of emotions in digital media. The sheer volume of such media renders it nigh impossible for humans to feasibly work with, and thus it falls to researchers to find viable methods for their automatic processing.

However, the same tremendous increase in processing power, storage volumes, and bandwidth that allows the users to generate unparalleled volumes of various media content has provided the means for the development of technologies for harnessing those. And so, the researchers continuously sought to employ the most advanced techniques to annotate the emotions in user-generated content. For emotion recognition provides a means to a range of ends, such as building a picture of a typical sentiment toward a public person or a phenomenon (Wang et al., 2020) or building an emotion-aware healthcare system (Ayata et al., 2020).

In the very beginning, emotion recognition mostly focused on speech, but as social networks, such as Facebook and Twitter, gained more and more users (WEB, a) and voice communication relatively withered, emotion recognition in text could not be ignored.

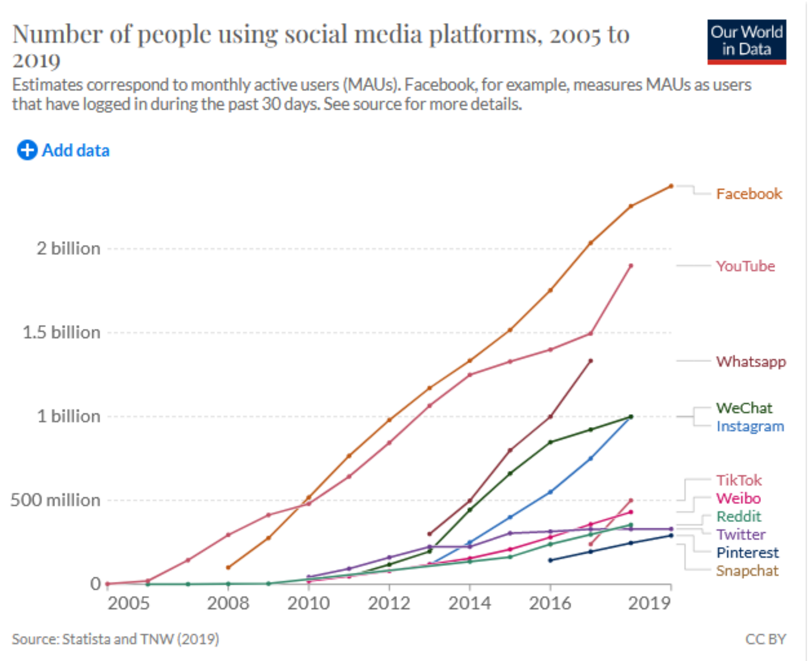


Fig. 1: Number of people using social media platforms from 2005 to 2019. Source: (WEB, a)

There is, however, a rather interesting facet of emotion recognition, that only comes to light when one observes the works researching the matter from a bird's eye view. Namely, that the absolute majority of researchers, at least, when it comes to authors presenting non-English datasets (Leonova, 2020) have based their annotation on Ekman's base emotion system (Ekman, 1992) that postulated the existence of six basic emotions, namely, anger, joy, sadness, surprise, disgust, and happiness. Some of the researchers amended the list, for example, removing disgust or adding fondness, but for the most part, the list was used as it is. We can speculate that the reason for this is its simplicity and popularity, but here comes the twist: even the revised version of the original model has removed the surprise from the list. Furthermore, a number of researches, such as (Gendron et al., 2014), have criticized the general concept of “basic” emotions and showed that the perception of emotions is not culturally universal. In much fewer cases, the researchers employed a two- or three-dimensional model that assigns valence, arousal, and dominance values to every emotion (Mehrabian, 1980), which provides a significantly more objective view on the emotion, alas, at the cost of simplicity and intuitiveness.

One of the shortcomings of the approach where the basic emotions are annotated, though, is that it overlooks such emotion as frustration, which, we surmise, deserves to be recognized among emotions being routinely experienced. However, as virtually any company nowadays strives to maintain a presence in major social networks, the frustration can serve as a measure of satisfaction or dissatisfaction experienced by users contacting those companies and thus be fairly helpful. However, nowadays there are only a handful of works that touch on the subject of frustration recognition, especially when talking about textual sources. From the other facet, the existing projects, dedicated to emotion recognition in text, with a few exceptions concentrate only on the words and their sequences. They use advanced tools such as n-grams and extensive vocabularies with calculated values of emotional “charge” for each word to determine the resulting emotional coloration of the text. However, this one takes into account lexical information, mainly ignoring the fact that, unlike the published text and its descendants, Twitter and other social network messages contain a wide range of non-lexical entities. Under “non-lexical” entities we understand any integral part of the text other than words. Those vary significantly in terms of sticking to the formal grammatical rules. Probably the most conventional are punctuation marks such as exclamation marks or ellipses, even though the number of those is not necessarily considered valid as far as rules of syntax are concerned. A bit less conventional, but already out of necessity recognized are omnipresent emojis, denoting happiness, merriment, or caring about another. And probably the least recognized are various ASCII and Unicode arts, such as ¯_(ツ)_/¯, bordering on nonce words and falling in and out of fashion. In addition, most of the works focusing on the lexical content are targeting English texts, while low-resource languages still struggle.

In our previous work (Leonova and Zuters, 2021) we tested our hypothesis that stipulated that the addition of features based on non-lexical means of expression (NLME) would improve the accuracy of frustration recognition and found that this hypothesis holds for the Latvian dataset.

In this work, we elaborate on our article (Leonova and Zuters, 2022) where we seek to demonstrate that the addition of NLME features derived from the Latvian dataset to the frustration recognition model is to a similar extent beneficial when applied to the English dataset. Here, we show that these achievements translate into a language-independent model, which can be speculated to extend into other languages within the same communicative culture; this, of course, calls for verification. Thus, the model is effectively language-independent (but not culture-independent) and can be employed for frustration recognition in English and, potentially, in any other language sharing the same cultural paradigm and, respectively, using NLME 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 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 NLME used by the bearers of the respective culture.

This paper has the following structure: at first, we present related works in Section 2; in the next section, we describe the datasets used for the experimentation. Section 4 presents the experimental setup and is followed by Section 5, which describes the presented model. Section 6 discusses the performance of the model and is followed by a review of possible future work and a conclusion.

2. Related Works

As we have mentioned before, emotion annotation in text and other media was a subject of keen interest for the last couple of decades. A system capable of emotion recognition in speech and synthesis of emotional content was presented as early as 1999 (Moriyama and Ozawa, 1999) and the very next year the researchers employed neural networks for the automation of this process (Nicholson et al., 2000). For a time, emotion recognition has concentrated on speech. The reason for that was that the textual content was not nearly as ubiquitous as it is nowadays and was ill-suited to defining the sentiments and disposition of the Internet population, as it mostly consisted of published works and periodic issues, or was highly specific, like themed discussion boards. Only five years later there started to appear researches aiming to derive emotions from text, first in multi-modal settings (Chuang and Wu, 2004), where emotions derived from the textual content of a speech would only play an auxiliary role. However, as time passed, there started to appear more and more human-generated content that could serve as a foundation for automatic emotion recognition based solely on text, and corresponding systems started to appear (Huang et al., 2005). While those earlier works were mostly keyword-based, as time passed, deep learning methods have started to be used in hybrid models in combination with the classic statistical methods (Seol et al., 2008) or alone (Ghazi et al., 2010). By now, these methods have mostly evolved into neural-network-based models in combination with extensive vocabularies, that list statistical weights of different words for various emotions, as well as word- and character-based n-gram features (Ameer et al., 2022).

When we speak about the annotation of emotions, however, especially in the context of low-resource languages with a small number of annotated corpora available for model training and calculating statistics, it can be seen that most authors still cling to annotating models based on Ekman’s six basic emotions: joy, anger, happiness, sadness, surprise, and disgust. They are sometimes used in their original form, for example, (Haryadi and Kusuma, 2019), or in a modified way, by adding or removing emotions from the list, with popular options being Plutchik’s (Plutchik, 2001) extension of the basic emotions by adding anticipation and trust as counterparts of surprise and disgust, for example, in (Semeraro et al., 2021) or addition of neutral emotion, such as in (Feng et al., 2021). Another popular variant is reducing the list of basic emotions by removing disgust (Araque et al., 2019) or both disgust and surprise (Mohammad et al., 2018), with more exotic variations ranging from replacing surprise with fondness (Yao et al., 2014) to recognizing 12, 15 and more finely discriminated emotions (Ameer et al., 2022). Less represented, but still universally recognized is using two- (Hofmann et al., 2021) or three-factor (Mohammad, 2018) models, which represent each emotion in the space of continual dimensions of valence, arousal, and (in three-factor models) dominance. As it can be seen, frustration is very rarely part of the deal, appearing in only a few works, like

(Hu et al., 2018), and is generally being understudied despite being potentially beneficial for such fields as customer support or service quality assessment.

As most of the state-of-art models employ impressive language-specific vocabularies and n-gram features for annotation, most researchers focus on English, it being the language with an enormous number of available resources available, and there are only a few resources targeting low-resource languages, such as Latvian, as, for example, (Gruzitis et al., 2018), or being language-independent, with tangentially relevant examples including language-independent sentiment analysis (Shakeel et al., 2019) or language-independent emotion recognition in speech (Singh et al., 2020).

Non-lexical means of expression (NLME), potentially universal within the range of languages sharing the same cultural context, have been studied sparingly and mostly in a dislocated manner, one or the other appearing in emotion annotation models. For example, usage of exclamation and question marks were used as a predictor by (Kirk et al., 2012) and the message length by (Hautasaari et al., 2019), but to the best of our knowledge, no systematic attempts were made before ours (Leonova, Zuters, 2021).

3. Datasets

For our experiments, we have effectively used two datasets, in English and in Latvian. Both datasets were comprised of Twitter conversations between users and customer support representatives answering using the company account. Each of those user messages were annotated by human annotators for the perceived level of the user's frustration expressed in this message on a scale of 0..4, where 0 would denote no frustration expressed and 4 would mean the maximal level of frustration. Using the results obtained by training the model on the Latvian dataset as a benchmark, we have tested our hypothesis using the English one, thus demonstrating that the model is language-independent. The English dataset represents the subset of the Kaggle Twitter messages dataset[1]. The authors have selected Twitter dialogs that fit the criteria and had them annotated for levels of frustration. It consists of 400 dialogs with 843 annotated user turns. The Latvian dataset was obtained by manually collecting dialogs from Twitter accounts of major telecommunication providers and contains 283 dialogs with 688 annotated user turns. User messages in both datasets were annotated by three independent annotators, and the median value was used as a resulting grade in further experiments. Both English and Latvian datasets, along with the code, are available on GitHub[2] and were described in detail in (Zuters and Leonova, 2020) and (Leonova and Zuters, 2021), respectively.

4. Non-Lexical Means of Expression

In our work, we use both lexical and non-lexical features as input to the model. As of the moment, the lexical means, i.e., words and their collocations, are relatively well-studied, and a range of elaborate tools, such as POS tagging, n-grams, and BERT is available for researchers, at least, when we speak about English or other high-resource languages, such as Chinese. To give an example, (Li et al., 2021) describe using BERT

[1] https://www.kaggle.com/thoughtvector/customer-support-on-twitter
[2] https://github.com/Lynx1981/dfrustration/tree/master/LatvianTweets

(Bidirectional Encoder Representations from Transformers) for text-based emotion detection.

On the other hand, there are but a few papers touching on non-lexical features of texts, even in high-resource languages. At most, isolated features, such as exclamation and question marks, hashtags (Mohammad and Kiritchenko, 2015), or the length of a text message were used along with the lexical features for the purposes of emotion annotation. In our research, we are using a relatively comprehensive set of non-lexical means of expression provided below, for predicting the level of frustration user messages. In order to determine the specific model input to be used in the experimentations, we have studied the corpus and identified a number of NLME features that could serve as potential predictors to the level of user frustration. For those, we calculated the correlation values with one another and with the assigned grade. The ones having at least a weak correlation with the annotated grade – in other words, with the level of frustration experienced by the user as perceived by the annotator, averaged, were used as input for our model. While the full correlation table and selection process is described in (Zuters and Leonova, 2020), we will provide a couple of examples. The message length was the best predictor, having the highest (positive) correlation of 0.44 with the annotated frustration level. At the same time, most of the features, such as the number of emojis, had a weak correlation of around ±0.1. The selected features are:

- Length of the message
- Number of exclamation marks in the message
- Number of question marks in the message
- Number of commas in the message
- Number of dots in the message
- Number of any of the three types of quotes in the message
- Number of uppercase words longer than four characters
- Number of positive emotions made up of typographical marks
- Number of negative emotions made up of typographical marks
- Presence of a picture in the message
- Number of "@" (effectively, other users' mentions in Twitter) in the message
- Number of repeating letter sequences in the message
- Presence of built-in smileys indiscriminate of valence in the message
- Number of digits in the message.

During the process of comparison of the results for the two languages, we encountered the need to adjust the measured characteristics to account for cultural differences and, accordingly, create a unified NLME feature. The first such case was letter repetition. While in the Latvian dataset the only letter that was repeated to imitate drawing out a vowel was "a", for example, in the word "draaaausmīgi lēni" ("increeedibly slow"), the most typical repetition in English was the one of letter "o", examples including "nooooo", "still slllllooooooow" and others. Thus, we have adjusted this feature to include the repetitions of all vowels.

The mentions of the Customer Protection Bureau of Latvia, as only specific for Latvian users, were replaced by the "@" character count, as in Twitter it is used to "tag" another user so that they would see themself mentioned. Of course, it can also be used in emails, but we speculate that the provision or indication of the email address by the user may likewise be characteristic for the cases where users either want a company representative to contact them or, on the opposite, mentions that he contacted the customer support via email.

Yet another feature, the number of digits in the message, was not used in the experiments before, but was identified as potentially beneficial during the review of the corpus with the goal of finding new potential features. We speculated that the number of digits may correlate with the perceived levels of frustration, as people tend to point at specific numbers as a representation of the problem that bothers them. However, the experiment did not justify our assumptions and the addition of the feature did not improve the resulting accuracy in any way.

The discussed features encode the non-lexical characteristics of the message and as such, construe the second part of the model input sequence, and the resulting model is described in Section 6, where the proposed model is discussed.

5. Experimental Framework for Frustration Intensity Prediction

In order to test our hypothesis and find the best meta-parameters, we have developed the following setup: we have constructed a neural network-based model that accepts a number of features as input. These features are constructed using a message from a user, addressed to a customer support representative. The first part of the input is constructed using the lexical features of the message by constructing a bag-of-words on the basis of interactively constructed vocabulary. The second part is constructed using NLME features described in the previous section. On the basis of this joint input, the system assigns to this message a grade representing the predicted level of frustration. This grade is then compared to the actual grade assigned to this message by (human) annotators, precisely, to the median value of the three. The median value was selected for an aggregated value in order to make use of averaging the grades while preserving the integer value thereof. The Python code for this model is available on GitHub along with the datasets used for training.

In our study, we have explored how the performance of the model was affected by variations in three different aspects. Our very first concern was tuning the model, thus we explored the performance resulting from using different parameters of the neural network itself. Neural networks, namely, multi-layered perceptrons with one hidden layer, were used as the technique to build our models upon as we have already successfully used them in the previous experiments and the main focus of the research is the frustration level analysis rather than a specific machine learning method used. After the meta-parameters have been established, we studied how preceding data processing, particularly, close-to-morphological segmentation of the message, affected the accuracy of predictions. Finally, we have focused on the main point of our research: how adding NLME features and their combinations to the input affected the performance of the model trained on the English dataset and how these effects compare to the results achieved on the Latvian dataset.

Thus, the overall experimentation was divided into two phases:

- Preparational phase of selecting experimental configurations empirically by conducting a few sets of experiments:
    - Establishing hyperparameters for neural networks,

  - Determining the effect of different preprocessing techniques and selecting conducive to model performance
  - Selecting the input configurations of lexical data for experimentation,
  - Establishing the best-performing set of NLME features.
- Main phase: running the experiments using the established model parameters on selected input configurations and comparing the results to the selected baselines.

First, we have established the optimal number of hidden neurons in the model, running the experiment with model configuration including 32, 64, 96 and 128 neurons. The results were that the model runs the best with 64 hidden neurons for both English and Latvian with 100 epochs being sufficient for our purposes; thus, all further experiments were conducted using this configuration.

The next was assessing the role of preprocessing in overall performance. For this purpose, the model was run with all available input parameters selected for use in this research at the stage of identifying and assessing NLME-based features, with the model configuration including 64 hidden neurons. This run was performed twice, first on the "untreated" dataset, and second, on the dataset, where all user messages were subjected to preceding segmentation with the GenSeg tool. After a comparison of the results, we had to conclude that the effects of data segmentation for the English dataset were ever so unexpectedly fully consistent with the ones obtained for Latvian: for both languages, message segmentation has improved the accuracy of predictions by slightly more than one percent.

After the model meta-parameters and the effect of segmentation were established, we turned to establishing the best-performing set of NLME features. To research those, we have run yet another series of experiments using various combinations of input features. To name the most prominent, we have used the single best feature, removal of underperforming features from the list, and all selected features, along with the bag-of-words set of features. Within this series of experiments, we did not in any way change the bag-of-words, as we have studied the effect of its different configurations for both English and Latvian datasets. So, for the experimentation, we used the bag-of-words consisting of one hundred words, most characteristic for the specific grade. Its construction is described in detail in the following section. Just like with the segmentation, we found that the results were consistent with the ones obtained on the Latvian dataset.

Were able to conclude that the behaviour of the model with different variants of input configuration is generally consistent over different languages, and the greater part of the accuracy is due to the four best predictors, while the complete removal of the features that produce no visible improvement when used in isolation, is disadvantageous to the resulting performance and leads to the slight decrease in accuracy.

As the performance of the original model on the English dataset was established, we have clarified whether the model adjusted for language and culture universality would not have the advantage of accuracy. For example, we used the number of PTAC (Consumer Rights Protection Bureau (of Latvia) mentions, which is inapplicable for English-speaking users; it was tentatively replaced with the count of the "@" symbol, which is used for mentions of other users, and the letter "a" repetition was complemented by letter "o" repetition, as it was the only repeated letter in English dataset, and other vowels, for potential encounters. The resulting slight increase in accuracy confirmed the soundness of this replacement.

Performance assessment was conducted by computing accuracy as a percent of correct frustration level predictions via leave-one-out cross-validation that consisted of training the model on all data except one entry and comparing the frustration in-tensity predicted for the one remaining (left out) entry, repeated for all the entries, averaged across fifteen runs. We use two points of reference: a neural model that only uses lexical features as input and the same model applied to the Latvian dataset.

6. Language-Independent Model to Measure Frustration Level

It stands to reason that to provide high-quality customer support and customer care it would be highly beneficial to be able to predict the level of frustration expressed in a customer's message, and to do so automatically, i.e., without human involvement. In modern times, when computational power is cheap and human labour is expensive, such automated processing can be used for the purposes of the triage and more optimal resource management. It can be even more useful in combination with profiling and other knowledge acquisition techniques. Such a model can come particularly handy if is language-independent, as in this case it can be used for low-resource languages, for which no extensive vocabularies with annotated emotions and n-grams exist, and would only require a relatively small dataset for training. Here we demonstrate that the proposed model, by utilizing the interactive vocabulary-building principles and language-independent (with the limitations, discussed above) features based on non-lexical means of expression, exhibits comparable performance in measuring the level of frustration for English and Latvian messages, addressed to a company customer support representatives via Twitter social network.

The model that we have developed is predicting the frustration level on a scale of 0 to 4, with zero denoting the absence of frustration and 4 representing the utmost level of frustration. The annotated dataset, however, contains two more additional markers. One is used to denote that it is impossible to determine the level of frustration from a message. This can be the case, for example, when a user answers a formal question, like stating their name or providing a phone number. The other indicates an incomprehensible message, for example, using other than the target language or suffering from technical issues, such as encoding. The messages with such grades are not used in model training. To be able to assign a value to a message, the model is taking advantage of three distinct features: interactive vocabulary construction, utilizing NLME-based features, and initial data processing.

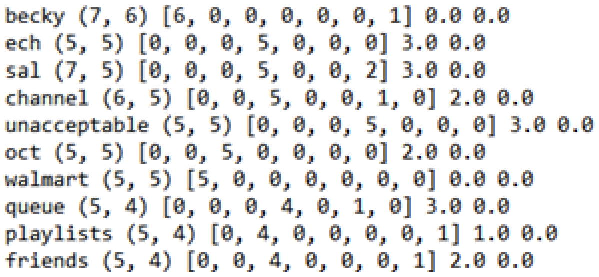
The first part of the features used as a model input is selected based on the lexical means of expression — namely, words, with the help of the interactively constructed vocabulary. The construction of such vocabulary was discussed in detail by (Zuters and Leonova, 2020), so we will only give a short outline here. To this end, during the training phase, all words in the training set are appraised for their predictive potential. To do so, for each word are calculated an average value of frustration for a message containing this word and the standard deviation of this annotated frustration grade among the messages in which this word appeared. Fig. 2 gives the excerpt from such vocabulary, constructed for the segmented English dataset. It is seen, that for each word in the dataset the following statistics are provided:

1) in round brackets: total number of messages, number of usable messages

2) In square brackets: number of messages for each value of frustration level: 0 through 4, message incomprehensible, impossible to establish the level of frustration,

3) after round brackets: aggregated value (median) of the annotated frustration level, its standard deviation across the messages.

For example, the very first line for the word segment “becky” gives the following information: the training set contained seven messages with this segment, and of those six were valid. Of those, six had the median annotated frustration value of 0 and one was deemed unfit for determining the level of frustration. The average annotated grade is 0.0 and its standard deviation was 0.0.



```
becky (7, 6) [6, 0, 0, 0, 0, 0, 1] 0.0 0.0
ech (5, 5) [0, 0, 0, 5, 0, 0, 0] 3.0 0.0
sal (7, 5) [0, 0, 0, 5, 0, 0, 2] 3.0 0.0
channel (6, 5) [0, 0, 5, 0, 0, 1, 0] 2.0 0.0
unacceptable (5, 5) [0, 0, 0, 5, 0, 0, 0] 3.0 0.0
oct (5, 5) [0, 0, 5, 0, 0, 0, 0] 2.0 0.0
walmart (5, 5) [5, 0, 0, 0, 0, 0, 0] 0.0 0.0
queue (5, 4) [0, 0, 0, 4, 0, 1, 0] 3.0 0.0
playlists (5, 4) [0, 4, 0, 0, 0, 0, 1] 1.0 0.0
friends (5, 4) [0, 0, 4, 0, 0, 0, 1] 2.0 0.0
```

Fig. 2: Top ten entries from the interactively constructed vocabulary.

After the statistics for every word from a training dataset have been calculated, the vocabulary is sorted by their predictive potential. We consider the best predictors the ones with the lowest standard deviation, where the standard deviation of zero would mean that this word only appears in messages with the specific annotated grade.

One hundred (as to why, see (Zuters, Leonova, 2020) top entries of the best predictor vocabulary are used to create a bag-of-words. This means that the lexical part of every message is coded as a sequence of one hundred binary values, where every binary value, 0 or 1, represents whether the corresponding word from the vocabulary was present in the message. Figure 3 illustrates this process.

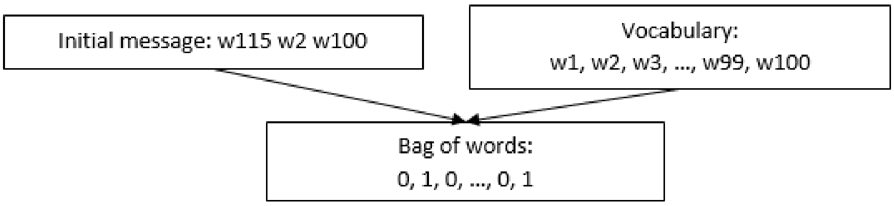


Fig. 3. The first part (bag of words) model input construction.

The second part of the input is constructed using NLME features, described in Section 4: Non-Lexical Means of Expression. Along with the bag-of-words they form the model input. Thus, the complete model input consists of 115 values: one hundred binary values for the presence of best predictor words in the message and fifteen numerical values, one for each NLME feature.

The last important part that we ought to present is input processing. As we've mentioned before, before constructing the dictionary, the input message is processed by subword segmentation (using the GenSeg tool (Zuters et al., 2019)), which helps to alleviate the noise resulting from different grammatical forms being used in the same context. We would like to mention that it does not affect the dictionary construction; independently of segmentation usage, the vocabulary is built based on the available lexemes. The only consequence of this would be that without the segmentation the vocabulary would be constructed for the complete words in as many forms are there are present in the training dataset. With the segmentation, however, the selfsame vocabulary will encompass both words and word segments, depending on the particular case.

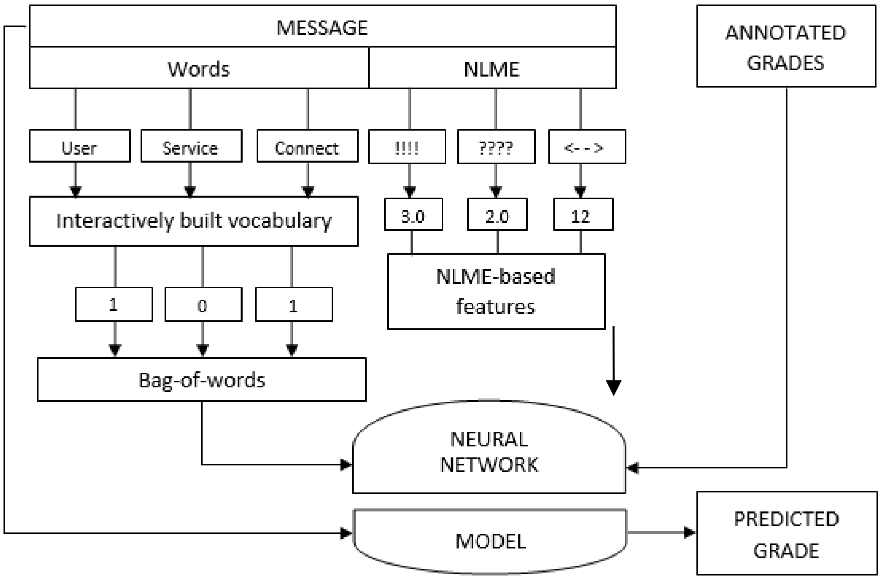


Fig. 4. Frustration level predicting model.

The detailed analysis of the experiment results is summarized in the following section.

7. Comparison of English and Latvian in Frustration Intensity Prediction

For both Latvian and English, we have produced a series of results for a number of model configurations, as described in section 4 "Experimental Framework". Specifically, the following questions were addressed:

- How the model augmented with NLME features perform in comparison with the reference model when applied to the English dataset and would the performance be dependent on these features similarly to the model applied to the Latvian dataset?

- Would the model's performance on the English dataset be affected by tuning the relevant NLME-based features to be more language-independent?
- Would the addition of yet another feature (the number of digits in the message) be beneficial for model performance?
- Would the segmentation improve the results similarly for both languages?

First of all, we gladly report that the two languages in their expression of emotions and perception thereof appear to be similar enough so that the identical set of features added to the model would improve its performance by 6pp from 41% to 47% accuracy which is slightly lower than 7pp accuracy (42% to 49%) improvement achieved for Latvian. In both cases, we are using as a baseline a bag-of-words model.

Not fully unexpectedly, we have found that the best performers were preserved: the most improvement was due to the set of four best features. Those are the length of the message, the number of exclamation marks in the message, the number of question marks in the message, and the number of dots in the message, accordingly. However, in a manner similar to Latvian, removing the underperforming features did not improve the result, but otherwise: without features, seemingly not contributing to the performance, the overall result was slightly lower — they jointly contributed around 0.5pp for both Latvian and English, totalling 48.3% for Latvian and 46.7% for English.

However, it has to be mentioned that two of the NLME features used for Latvian were country- and language-specific, as we have used as an indication a number of PTAC (Customer Rights Protection Bureau) mentions in a message, which obviously does not apply to English tweets, as well as the repetitions of the letter “a”, that was characteristic for Latvian, but not English language; there, the repetition of “o”, was more widespread. In order to adjust the model to suit the dataset in English and to be more language-independent in general, we have replaced it with a feature that calculates the number of “@” symbols repeated in the message that is universally used for mentions in various social networks and counted in the repetition of all vowels. This has improved the accuracy by 0.6pp, resulting in a total of 48.2% (7.3pp of improvement compared with the baseline). At the same time, this universalization with changing the mentions of Customer Protection Bureau into universal mention (and, to less extent, email) count also proved to be beneficial for Latvian, improving the accuracy score by 0.3pp.

The addition of the number of digits in the message was found not to be beneficial either for English or Latvian and did not result in any increase in performance.

The results are summarized in Tables 1 and 2.

Table 1. Frustration prediction accuracies (%) for various proposed models.
C1 - NLME model with all features, C1* – features adjusted (**both** with and without digit count), C2 - NLME model without subpar features, C3 - NLME model with all features and no segmentation, C4 - NLME model with a single best feature, RM – reference model

Model	C1	C1*	C2	C3	C4	RM
Latvian	48.8	49.1	48.4	47.5	46.9	42.1
English	47.2	48.2	46.7	45.9	43.7	40.9

Table 2. Frustration prediction improvements (pp) against the reference model for various proposed models.
C1 - NLME model with all features, C1* – features adjusted (both with and without digit count), C2 - NLME model without subpar features, C3 - NLME model with all features and no segmentation, C4 - NLME model with a single best feature.

Model	C1	C1*	C2	C3	C4
Latvian	6.7	7.0	6.3	5.4	4.8
English	6.3	7.3	5.8	5.0	2.8

What has come as a surprise, though, was the role of segmentation in the overall result. While for Latvian, being a synthetical language with a lot of flexions and grammatical forms, the slight improvement of 1.25pp achieved by segmentation of the source data, was to a certain degree expected, the same result of 1.25pp achieved for mostly analytical English was not. We can only speculate that accounting for noun singular and plural together, as well as stringing together different verb forms such as third person and participle provide for that difference.

Summarizing our findings, we can tell that the performance improvement resulting from extending the model with NLME features and data segmentation appears to be transferable to another language, namely English, from Latvian, for which this set of features was initially developed, to the full extent. That is, the extension of the bag-of-words model improves the results by 6pp or 7pp, of which the increase of 1.25pp is achieved due to the subword segmentation of the data. The removal of underperforming features causes a decline in resulting accuracy. The best results are acquired using 64 hidden neurons over 100 epochs.

8. Future Works

In this study, we have researched whether the presumably language-independent model, originally developed using the Latvian dataset, would be applicable to the English data. As we have demonstrated, it is indeed working as intended. However, Latvian and English both belong to the Indo-European language family, thus raising a question: Does the applicability of the proposed model cross the border of the language family, and whether, being sufficiently augmented, it could be applied to non-alphabetic languages? In the future, we would like to explore those possibilities. In addition, we want to research the extensions of the NLME set, should this prove possible, by studying the relevant data sources and possibly involving professional linguists for their insights.

9. Conclusion

The development of social networks and the explosive growth of user-generated content made it nearly impossible for basically any company or person of notice to keep afloat without employing social media to keep in contact with the target audience, for both sharing information and receiving feedback. Companies nowadays routinely use social networks to launch web-oriented campaigns and react to users' mentions and messages. However, due to the enormous volume of content, it might be beneficial to employ one or the other automation technique in order to stay informed of relevant trends, tendencies, and sentiments. Emotion annotation plays a vital part in such methods and systems and thus keeps being the object of keen interest of numerous modern researchers. The existing works are mostly focusing on annotating basic emotions, while frustration is underrepresented despite being of practical interest in such areas as customer support, customer satisfaction, and alike.

In our previous works, we have presented a neural network-based model that targeted measuring the level of frustration on a scale of 0 to 4 in the Twitter messages with interactively built vocabulary (Zuters and Leonova, 2020) based on best predictor words and showed how non-lexical means of expression and segmentation can improve the predictions (Leonova and Zuters, 2021) on the material of the annotated dataset in Latvian.

In this work, we observed the performance of the model developed on the material of the Latvian dataset and the role of input segmentation when applied to the English dataset. For those purposes, we have used the manually annotated dataset consisting of user dialogues with customer support representatives on Twitter. We have demonstrated that input data processing as well as the features initially developed on the Latvian material are providing a similar increase in accuracy, even more so after the adjustment of features for a higher extent of language-independence. In addition, we tested the addition of a new feature, the number of digits in a message, which proved to not improve the accuracy for both datasets. As a baseline for comparison, we are using the accuracy, achieved by the model without the employment of data processing methods or NLME-based features. The baseline is approximately 42% for Latvian and 41% for English. The model, employing both NLME and data processing, achieves an accuracy of 47% for English and 49% for Latvian, which amounts to approximately a 6pp and 7pp increase in accuracy. However, provided the features are adjusted in accordance with English data, the resulting accuracy achieved on the English dataset comprises 48%, which is 7pp over the reference model and is similar to the results achieved for Latvian.

References

Ameer, I., Sidorov, G., Gómez-Adorno, H., Nawab, R.M.A. (2022). Multi-label Emotion Classification on Code-Mixed Text: Data and Methods. *IEEE Access*.

Araque, O., Gatti, L., Staiano, J., Guerini, M. (2019). Depechemood++: A Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques. *IEEE transactions on affective computing*.

Ayata, D., Yaslan, Y., 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.

Chuang, Z.J. and Wu, C.H. (2004). Multi-modal emotion recognition from speech and text. In International Journal of Computational Linguistics & Chinese Language Processing, **9**(2), August 2004: Special Issue on New Trends of Speech and Language Processing (pp. 45-62).

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

Feng, S., Wei, J., Wang, D., Yang, X., Yang, Z., Zhang, Y., Yu, G. (2021). SINN: A speaker influence aware neural network model for emotion detection in conversations. *World Wide Web*, **24**(6), pp.2019-2048.

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

Ghazi, D., Inkpen, D., Szpakowicz, S. (2010). Hierarchical approach to emotion recognition and classification in texts. In *Canadian Conference on Artificial Intelligence* (pp. 40-50). Springer, Berlin, Heidelberg.

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

Haryadi, D., Kusuma, G.P. (2019). Emotion detection in text using nested long short-term memory. *11480 (IJACSA) International Journal of Advanced Computer Science and Applications*, **10**(6).

Hautasaari, A., Yamashita, N., Gao, G. (2019). How non-native English speakers perceive the emotional valence of messages in text-based computer-mediated communication. *Discourse Processes* **56**(1). pp. 24-40.

Hofmann, J., Troiano, E., Klinger, R. (2021). Emotion-aware, emotion-agnostic, or automatic: Corpus creation strategies to obtain cognitive event appraisal annotations. *arXiv preprint arXiv:2102.12858*.

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

Huang, X., Yang, Y., Zhou, C. (2005). Emotional metaphors for emotion recognition in Chinese text. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 319-325). Springer, Berlin, Heidelberg.

Kirk, R., Roach, M.A., Johnson, J., Guthrie, J., Harabagiu, S.M. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. In Lrec, vol. 12, pp. 3806-3813. 2012.

Leonova, V. (2020). Review of non-English corpora annotated for emotion classification in text. In *International Baltic Conference on Databases and Information Systems* (pp. 96-108).

Leonova, V., Zuters, J. (2021). 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., 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) Forum.* 2022

Li, M., Chen, L., Zhao, J., Li, Q. (2021). Sentiment analysis of Chinese stock reviews based on BERT model. *Applied Intelligence*, **51**(7), pp.5016-5024.

Mehrabian, A. (1980). Basic Dimensions for A General Psychological Theory. pp. 39–53.

Mohammad, S. M. (2018). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1-17).

Mohammad, S.M., Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, **31**(2), pp.301-326.

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

Nicholson, J., Takahashi, K., Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural computing & applications*, **9**(4), pp.290-296.

Plutchik R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Shakeel, M.H., Faizullah, S., Alghamidi, T., Khan, I. (2020). Language independent sentiment analysis. In *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)* (pp. 1-5). IEEE.

Semeraro, A., Vilella, S., Ruffo, G. (2021). PyPlutchik: Visualising and comparing emotion-annotated corpora. *Plos one*, **16**(9), p.e0256503.

Seol, Y.S., Kim, D.J., Kim, H.W. (2008). Emotion recognition from text using knowledge-based ANN. In *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications* (pp. 1569-1572).

Singh, R., Puri, H., Aggarwal, N., Gupta, V. (2020). An efficient language-independent acoustic emotion classification system. *Arabian Journal for Science and Engineering*, **45**(4), pp.3111-3121.

Wang, S., Schraagen, M., Sang, E.T.K., Dastani, M. (2020). Public sentiment on governmental COVID-19 measures in Dutch social media. openreview.net

WEB (a). https://ourworldindata.org/grapher/users-by-social-media-platform as of 2022-10-11

Whiten, A., 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.

Yao, Y., Wang, S., Xu, R., Liu, B., Gui, L., Lu, Q., Wang, X. (2014). The construction of an emotion annotated corpus on microblog text. *Journal of Chinese Information Processing*, **28**(5), pp.83-91.

Zuters, J., 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., Strazds, G., Leonova, V. (2019). Morphology-Inspired Word Segmentation for Neural Machine Translation. In *Databases and Information Systems X*, pp. 225-239. IOS Press,.

Received October 31, 2022, accepted November 29, 2022