Emotion analysis in socially unacceptable discourse

Jasmin FRANZA
Faculty of Arts, University of Ljubljana

Bojan EVKOSKI
Jožef Stefan International Postgraduate School; Jožef Stefan Institute

Darja FIŠER
Faculty of Arts, University of Ljubljana; Jožef Stefan Institute;
Institute of Contemporary History

Texts often express the writer's emotional state, and it was shown that emotion information has potential for hate speech detection and analysis. In this work, we present a methodology for quantitative analysis of emotion in text. We define a simple, yet effective metric for an overall emotional charge of text based on the NRC Emotion Lexicon and Plutchik's eight basic emotions. Using this methodology, we investigate the emotional charge of content with socially unacceptable discourse (SUD), as a distinct and potentially harmful type of text which is spreading on social media. We experiment with the proposed method on a corpus of Facebook comments, resulting in four datasets in two languages, namely English and Slovene, and two discussion topics, LGBT+ rights, and the European Migrants crisis. We reveal that SUD content is significantly more emotional than non-SUD comments. Moreover, we show differences in the expression of emotions depending on the language, topic, and target of the comments. Finally, to underpin the findings of the quantitative

Franza, J., Evkoski, B., Fišer, D.: Emotion analysis in socially unacceptable discourse. Slovenščina 2.0, 10(1): 1–22.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.1.1-22

investigation of emotions, we perform a qualitative analysis of the corpus, exploring in more detail the most frequent emotional words of each emotion, for all four datasets. The qualitative analysis shows that the source of emotions in SUD texts heavily depends on the topic of discussion, with substantial overlaps between languages.

Keywords: emotions, socially unacceptable discourse (SUD), hate speech, social media, corpora

1 Introduction

Emotions are a key component of human behaviour and communication. Most often, we use language to manifest, transmit and explain emotions. Meanwhile, the continuously increasing popularity of social media produces unprecedented amounts of user-generated content from people all around the world and in all languages. Oftentimes, this content (posts, comments, descriptions etc.) includes words that reveal the scope of emotions the author tries to unveil while evoking specific emotions from the reader as well. The social media era has also introduced very open outbursts of socially unacceptable discourse (SUD), such as hate, discriminatory, offensive or threatening speech. This has given rise to the necessity of analysing SUD communication practices in order to better understand and effectively tackle them. Here we dive into the field of Emotion Recognition (ER), which aims to recognize and categorize verbalized emotions in texts. By doing so, we hope to understand what outlines SUD content through the viewpoint of emotions.

In this paper, we introduce a novel, yet simple method to analyse emotions in text by utilizing the NRC Emotion Lexicon (Mohammad and Turney, 2010). A metric which we name Emotional Charge (E_C), calculates the overall emotion intensity of a comment. We utilize our approach on social media content by answering how emotions depend regarding the language, topic and most importantly, its SUD contribution.

For that purpose, we focus on emotions expressed in socially unacceptable Facebook comments in Slovene and English on the topics of the European migrant crisis (hereinafter referred to as Migrants) and LGBT+ rights (hereinafter referred to as LGBT+) from the FRENK data-

set (Ljubešić et al., 2021), as it is a uniquely carefully annotated multilingual dataset on SUD content which covers two topics. We perform a quantitative analysis for both languages and topics, taking into account the degree of emotional charge in each comment and the representation of individual categories of emotions by using the NRC emotion lexicon, which organizes words into one of the eight basic emotions by Plutchik (1980). To complement the quantitative approach, a qualitative analysis of the emotional words in SUD comments is added, enabling a more thorough understanding of the emotional charge findings.

The two main research questions covered in this paper are:

- Does SUD content differ from non-SUD content in the expression of emotions?
- Does the emotional footprint of SUD comments differ depending on the topic and target they address?

The paper is organised as follows. Section 2 gives an overview of the background and related work; Section 3 focuses on the description of the dataset used and describes the methods for calculating the emotional charge of comments. Subsequently, Section 4 presents the analysis on the emotional landscape of SUD, both from a statistical point of view and from a qualitative angle giving a deeper look at the emotional lexicon connected to SUD. Finally, Section 5 concludes the paper with a discussion and ideas for future work.

2 Background and related work

In the past decade, there has been an increase in research in the field of automatic detection of emotions in user-generated content (Alm et al., 2005; Al-Saqqa et al., 2018). However, although SUD has been intensively analysed in various disciplines and methodological frameworks, approaches to SUD via emotion analysis has so far received little attention (Gitari et al., 2015; Martins et al., 2018). This article presents an approach to comprehensively analyse SUD with the help of emotion lexica, as Markov et al. (2021) showed that emotion-based features provide useful cues for its automatic detection. In this section, we present the theoretical underpinnings for the analytical part of our study.

2.1 Emotions

In psychology, there is no general unanimity on the definition of emotions and their number. Research mostly focuses on two approaches to the representation of emotions, namely the category model and the dimensional model (Scherer, 2005). In the category model, emotions are presented as sets of different basic emotional states (e.g., JOY, ANGER) where basic emotions are understood as those that appear in very early childhood development and their expression and recognition are culturally independent. In the dimensional approach, emotions are presented in the space where each emotion occupies its place in an emotion dimension (e.g., value dimension: positive-negative axis, strength axis: high-low; Russell, 1980). The categorical approach is more widespread in computational linguistics than the dimensional one (Aman and Szpakowicz, 2007; Ghazi, 2016) because it is more intuitive and easier to apply, especially in computational models, which is why we adopt it in the study presented in this paper. We use the categorization into 8 basic emotions according to Plutchik (1980), namely JOY, SADNESS, ANGER, FEAR, TRUST, DISGUST, SURPRISE and ANTICIPATION, as they represent the basic and prototypical emotions, with the combination of which we can build more complex ones, e.g., LOVE, AWE, CONTEMPT. This model is also called Plutchik's wheel of emotions, as each fundamental emotion also has its opposite emotion (e.g., JOY - SADNESS, FEAR – ANGER; Plutchik, 2001; see Figure 1). Their organisation is based on the physiological purpose of each.

Martins et al. (2018) show that the most critical emotions to identify hate speech are the negative ones – ANGER, DISGUST, FEAR AND SADNESS as they occur in 2/3 of hate speech texts, while they claim SURPRISE can be interpreted as a neutral emotion in hate speech. On the other hand, ANTICIPATION, JOY and TRUST can be classified in the positive emotions group.

2.2 Emotion Recognition from Text

Emotion recognition from text can be divided into two groups: the earlier approaches, based on lexical datasets (Mohammad and Turney, 2010), and the latter ones, based on annotated training corpora (Aman and Szpakowicz, 2007; Canales et al., 2019). In the corpus approach,

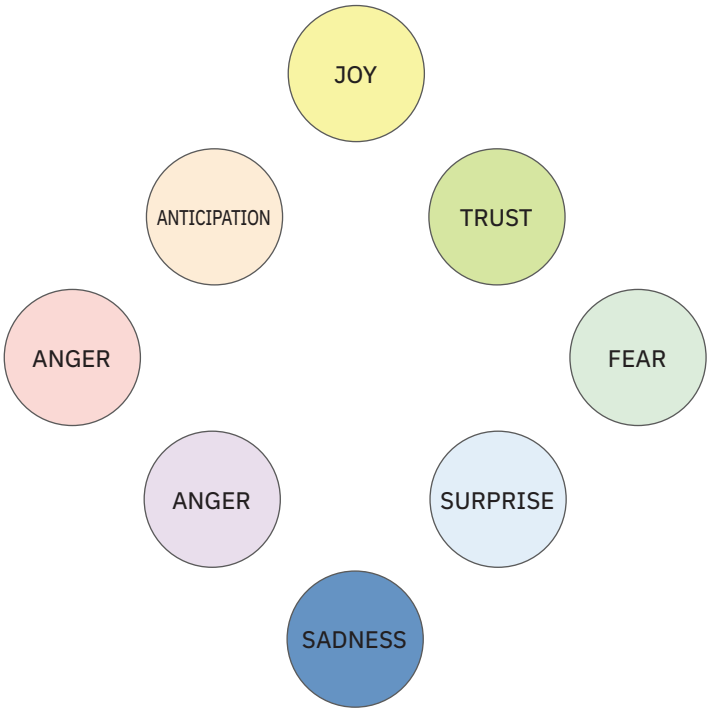

Figure 1: PLUTCHIK'S WHEEL OF EMOTIONS. It shows 8 basic emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust.

machine learning methods based on pre-annotated texts are employed to develop models for annotating new texts, while the lexical approach for identifying emotions in texts uses an external set of vocabulary with emotion tags. Due to the greater universality and adaptability to different domains and genres, we follow the latter paradigm. Additionally, previous research confirms the adequacy of the lexical approach. Mohammad and Yang (2011) have successfully used the NRC Emotion Lexicon to identify predominant emotions in love letters, hate emails, and suicide records. They were mainly interested in the difference between the linguistic expression of emotions in men and women. A similar approach with the help of the SentiWordNet lexical database, which contains strings of synonyms with assigned sentiment tags, was successfully used by Denecke (2008) on machine-translated texts to predict sentiment.

2.3 Socially Unacceptable Discourse (SUD)

Hate speech is a widespread phenomenon that attracts many researchers from diverse areas. However, the term is usually used in a very narrow, legally defined sense in the literature, which is why we adopt the term Socially Unacceptable Discourse (SUD), comprising all forms of hateful, discriminatory, offensive, violent or threatening speech (Fišer et al., 2017). A significant part of contemporary SUD research takes place within critical discourse analysis in combination with corpus linguistics (cf. Brindle, 2016; Knoblock, 2017), intending to identify and analyse SUD and its evolution. Assimakopoulos et al. (2017) present several European research projects on SUD, in which the analysis of online content is predominant. They point out that EU legislation alone is not enough to solve the spread of online hate and improve its understanding, as SUD can manifest itself in many subtle ways, such as stereotyping and categorization, patriotism, metaphorical expression, sarcasm, allusions etc., which makes comprehensive linguistic approaches extremely important for better awareness of the issue. A deeper understanding of SUD would mean principally better prevention and identification.

In Slovenia, the most valuable resource of SUD data is the manually annotated FRENK corpus of Facebook comments (Ljubešić et al., 2019; Ljubešić et al., 2021; see Section 3). Vehovar et al. (2020) show that about half of all the comments appearing in the FRENK dataset were identified as SUD: the share is significantly higher for the topic of Migrants (58%) than for the topic of LGBT+ (48%). The dataset was also analysed from the linguistic point of view, revealing SUD has a different lexical footprint (Franza and Fišer, 2019) and showing SUD comments are less standard than non-SUD comments with also a lower frequency of emoticons/emojis and punctuation (Pahor de Maiti et al., 2019).

3 Dataset and methodology

In this section, we describe in detail how the FRENK dataset, which is also used in this paper, was constructed and how it was processed for the purposes of this analysis. Next, we present the NRC Emotion Lexicon and the emotion labels it contains. Based on these, the emotional

charge of each comment in the FRENK dataset is calculated, which is presented in the final subsection.

3.1 FRENK Corpus

The FRENK corpus (Ljubešič et al., 2019; Ljubešić et al., 2021)[1] was collected from Facebook pages of three mainstream news media outlets for each examined language, including Slovene and English. It covers two topics, the EU Migrants crisis and the LGBT+ rights, and was enriched with manual annotations of the comments (Ljubešič et al., 2019). The Slovene part of the corpus contains 30 posts with 6545 comments for Migrants, and 93 posts with 4571 comments for LGBT+. The English part of the corpus consists of 16 posts with 5855 comments for Migrants, and 14 posts with 5906 comments for LGBT+. Additionally, comments were annotated for the type of SUD they produce (acceptable, background-violence, background-offence, other-threat, other-offence and unacceptable), as well as a categorization of the people being the target of the comment (Migrants, members of the LGBT+ community, persons related to Migrants or LGBT+, journalists or media, fellow commenter, other).

The dataset is linguistically processed with the CLASSLA pipeline for Slovene (Ljubešić, 2019, 2020) and Stanza for English (Peng Qi et al., 2020) on the levels of tokenization and sentence splitting, PoS-tagging and lemmatization. Therefore, we were able to annotate the lemmatized English and Slovene datasets with the NRC Emotion Lexicon for the corresponding language, which resulted in a bilingual and comparable emotion-labelled dataset of SUD Facebook comments that we analyse in the remainder of this paper.

3.2 Emotion Annotation

To identify emotions, we used the NRC Emotion Lexicon. The lexicon contains all words from Roget's Thesaurus that appear more than 120,000 times in Google's n-gram corpus, resulting in 14,200 entries.

1 The FRENK corpus, besides its Slovene and English parts, was created also for Croatian, French and Dutch, http://nl.ijs.si/frenk/english. The Dutch version was created within the LiLaH project: https://lilah.eu.

Each word in the lexicon has a label for its polarity (positive, negative) and for Plutchik's 8 basic emotions (ANGER, ANTICIPATION, DISGUST, FEAR, JOY, SADNESS, SURPRISE, TRUST). It was annotated manually using the crowdsourcing platform Amazon Mechanical Turk. The lexicon was originally created for English, and was later also automatically translated into 105 languages, including Slovene, with the help of Google Translate (2017). We have performed manual post-editing of the machine-translated lexicon (Daelemans et al., 2020). Examples of the translated lexicon along with the emotion labels can be found in Table 1.

Table 1: *Examples of emotion annotation in the NRC Emotion Lexicon*

English	abandoned	happiness	wise	ghost	refugee
Slovene translation	opuščen, zapuščen, prekinjen, zavržen	sreča, veselje	moder	duh, prikazen	begunec
ANGER	Yes	No	No	No	No
ANTICIPATION	No	Yes	No	No	No
DISGUST	No	No	No	No	No
FEAR	Yes	No	No	Yes	No
JOY	No	Yes	No	No	No
SADNESS	Yes	No	No	No	Yes
SURPRISE	No	No	No	No	No
TRUST	No	Yes	No	No	No

Note. For each English entry, there is a manually post-edited machine translation in Slovene and annotations for each of the 8 basic emotions.

3.3 Lexicon Limitations

The lexical approach is an efficient method to tackle emotion recognition from text (cf. 2.2). It is essential to work with datasets that are carefully prepared and verified to have reliable results. Our approach in this paper tests this method and achieves interesting outcomes. Nonetheless, it is important to also state the limitations of this specific emotion lexicon, the NRC emotion lexicon (Mohammad and Turney, 2010). We identified two main issues, namely the presence of biases and questionable emotion labelling.

Our work focuses on SUD, and it is important to point out that the lexicon has non-neutral annotations for the two topics we are dealing with, which can be linked to the lack of control and documentation about who the annotators were in the first place as the lexicon is the result of crowdsourcing. For example, *immigrant* is annotated with FEAR, *fugitive* with FEAR, ANGER, DISGUST, SADNESS and TRUST, *lesbian* with DISGUST and SADNESS. It is possible to note that there are some prejudices in these labels and it could be problematic as our work aims to fight against biases. Moreover, some labels appear to be ambiguous. For example, *nurture* is annotated with ANGER, ANTICIPATION, DISGUST, FEAR, JOY and TRUST, which suggests contradictory emotions together and does not give an insightful perspective of the word.

There have been many attempts to create an emotion lexicon, but the NRC emotion lexicon attracted the most attention due to its availability, size, and its choice of Plutchik's expressive eight-class emotion model (Zad et al., 2021). This is also the reason why we decided to use it, but we will take into account the potentially problematic labels in our interpretation of the results and we will complement the analysis with a qualitative study to check for potentially problematic consequences of using the lexicon. There have been also several attempts to improve the NRC emotion lexicon (cf. Zad et al., 2021), which should be further explored in the future.

3.4 Lexicon Coverage

Table 2 shows statistics regarding the NRC lexicon coverage of our dataset, for each of the subsets. Lexicon coverage has been calculated as the percentage of unique emotionally eligible words found in the lexicon, which means not all of them are labelled with emotion tags. The English language subsets contain around 5000 to 9000 unique words, with the NRC lexicon coverage of 20%. Meanwhile, the Slovene subsets, although of similar size, contain more unique words, with around 6000 to 12,000, depending on the dataset. Expectedly, since the Slovene NRC lexicon is the result of the machine-translated English lexicon and has a generally higher number of unique words, the NRC lexicon coverage of the Slovene dataset is a bit lower (around 16%), with small differences depending on the subset. Manual examinations have shown that there is a small

number of false positives and a higher number of false negatives, implying that the lexicon should be further improved. A random (subjective) sample evaluation of 100 English and 100 Slovene comments on the performance of the lexicon revealed the following:

- English NRC lexicon: Precision – 0.96; Recall – 0.65
- Slovene NRC lexicon: Precision – 0.91; Recall – 0.64

The low recall for both English and Slovene shows that the lexicon fails to recognize a large portion of the emotional words present in the comments, which is expected as we focus on a very specific kind of discourse on social media with specific characteristics on a very narrow topic. It is possible to find an explanation for the low recall also in the false negative emotionally eligible words (emotional, but not covered by the lexicon), as for example *shootings*, *frightened*. Moreover, SUD comments exhibit a peculiar tendency towards nonstandard features (Pahor de Maiti et al., 2019), which compromises emotional words recognition, for example *strelat* instead of *streljati* (eng. *to shoot*). Additionally, the evaluation indicates a lower precision of the Slovene lexicon, which could be possibly explained because of more false positives (not emotional, but included in the lexicon), which are mainly due to polysemy and non-canonically spelled words. For example, the Slovene lexicon contains the adjective *sam* (eng. *alone*), but in the comments it is used as an adverb (meaning *just*), which should not be an emotional word.

Table 2: *Statistics regarding the NRC Lexicon coverage of our dataset, divided per topic, language and SUD/non-SUD comments*

Language	Topic	SUD/ non-SUD	Comments	Unique words	Emotionally eligible words (nouns, verbs, adjectives and adverbs)	Lexicon coverage
English	Migrants	Non-SUD	2964	8401	5291	1046 (20%)
English	Migrants	SUD	2867	9323	6818	1323 (19%)
English	LGBT	Non-SUD	1777	8514	5374	1124 (21%)
English	LGBT	SUD	4080	5622	4297	977 (23%)
Slovene	Migrants	Non-SUD	2646	8401	5889	863 (15%)
Slovene	Migrants	SUD	3795	12486	10020	1325 (13%)
Slovene	LGBT	Non-SUD	1855	6199	4745	878 (19%)
Slovene	LGBT	SUD	2606	10108	8392	1329 (16%)

3.5 Calculating Emotional Charge

The final stage is to use the lemmatized comments and the lexicon to calculate a metric that defines the overall emotion intensity. We introduce this metric in order to be able to compare comments not just on the level of a specific emotion, but also have a universal comparison which includes all, answering the questions posed in the Introduction.

We define Emotional Charge (E_C) as follows: let W be the list of all nouns, verbs, adjectives and adverbs in one comment (as the emotionally eligible word functions). Then, let W_E be the list of all words in W which are labelled as emotional by the emotion lexicon. We define emotional charge E_C of a comment as follows:

$$E_C = \frac{|W_E|}{|W|}$$

To put it simply, Emotional Charge (E_C) calculates the portion of emotional words labelled by the lexicon in the total number of emotionally eligible words. Note that W and W_E are defined as lists and not as sets, thus an emotional word being present twice in a comment is also counted twice in the total score.

Using the emotional charge of each comment, we were able to get a sampling distribution of emotional charge for the desired group of comments (e.g., SUD vs. non-SUD, Slovene vs. English, Migrants vs. LGBT+). Figure 2 shows an example of the procedure for calculating the emotional charge.

Taking into account only word types that can contain emotion (nouns, verbs, adjectives and adverbs) as well as using the emotional charge formula that normalizes comment length makes the emotional charge scores more robust. Yet, this way of calculating emotional charge introduces many “non-emotional” and “highly” emotional short comments, where the emotional charge is 0 or 1 respectively, based on only a few words. Thus, we made a pragmatic decision of excluding comments with less than three words from the rest of our analysis.

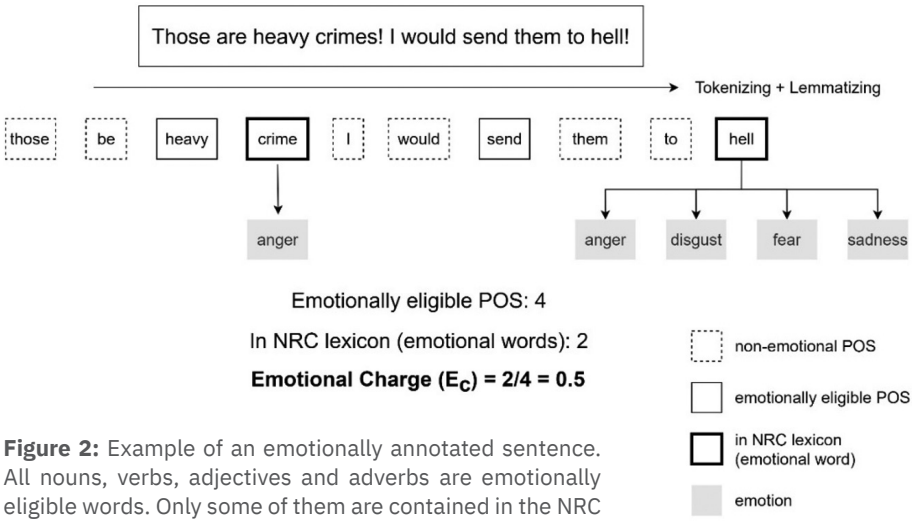

Figure 2: Example of an emotionally annotated sentence. All nouns, verbs, adjectives and adverbs are emotionally eligible words. Only some of them are contained in the NRC Emotion Lexicon. The ones in the Lexicon are counted in the emotional charge of the sentence.

4 Results

In this section, we present our research findings using the emotional charge of SUD comments for both Slovene and English language.

4.1 Emotional Charge Analysis

SUD is more emotional than non-SUD content. Here, we check whether emotion annotation is informative for differentiating between SUD and non-SUD comments by comparing their emotional charge. Once we calculated the distribution of emotional charge for each of the groups, we applied the Kolmogorov-Smirnov two-sample test (Pratt and Gibbons, 1981), which showed a statistical difference between SUD and non-SUD across all four combinations of language and topic (Migrants English $p=3\times10^{-7}$; $d=0.18$, LGBT+ English $p=3\times10^{-8}$; $d=0.23$, Migrants Slovene $p=1\times10^{-10}$; $d=0.18$ and LGBT+ Slovene $p=3\times10^{-4}$; $d=0.12$). The effect size d according to Cohen's formula (Cohen, 1988) is considered small to medium (depending on the combination). Figure 3 shows the distribution mean and deviation of all four combinations. Thus, we conclude that a specific analysis on SUD content could indeed be informative as the data showed that these comments are significantly more emotionally charged than non-SUD.

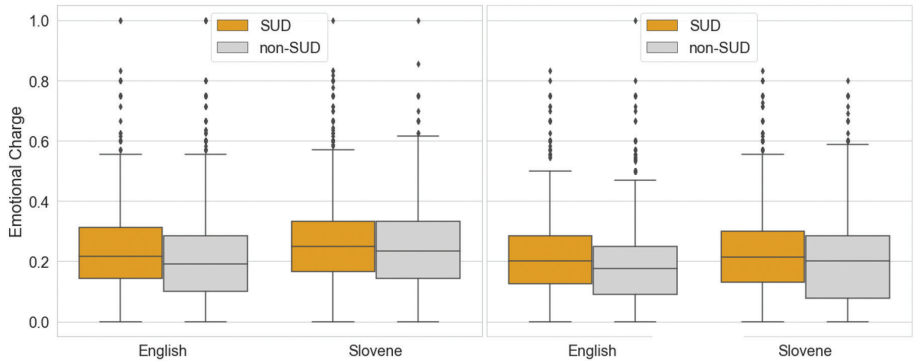

Figure 3: Comparison of emotional charge between languages and topics of SUD comments and non-SUD comments. The figure shows distributions (rectangles) and variance (lines), SUD comments are significantly more emotionally charged than non-SUD ones.

Topics differ in emotional charge – LGBT+ evokes more emotions than the Migrants topic. After confirming a higher emotional charge in SUD comments, we analysed whether one topic attracts more emotional charge than the other. Figure 4 shows, side-by-side, the distributions of the sets we compare (LGBT+ vs. Migrants). The Kolmogorov-Smirnov test suggests that there is a statistical difference in the emotional charge between Migrants and LGBT+, as both in English ($p=3\times10^{-9}$) and Slovene ($p=2\times10^{-11}$) comments, the LGBT+ topic carries a higher emotional charge. According to Cohen's coefficient, in English, the effect size is medium ($d=0.209$) while in Slovene it is small (d=0.183).

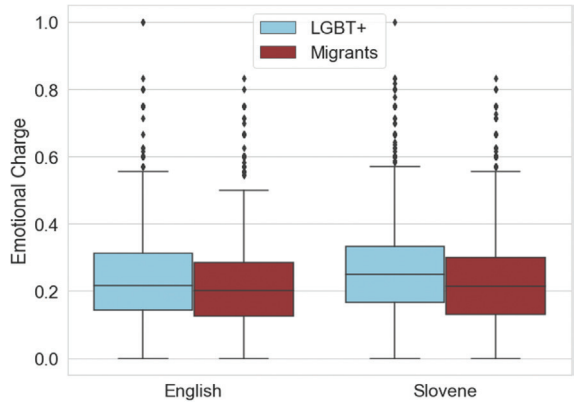

Figure 4: Difference of SUD Emotional charge between the LGBT+ and migrant topics in English and Slovene. The figure shows distributions (rectangles) and variance (lines), resulting in LGBT+ comments being more emotionally charged.

Comments are more emotional when targeted at Migrants/LGBT+. One of the metadata information of the FRENK dataset is the target of the comment, or in other words, who the comment is directed at. We compared *"the commenter"* and *"the target – migrant/LGBT+ person"* which are the two most frequent targets in the FRENK dataset (see Section 3.1). The comments targeted at *migrants/LGBT+ persons* are explicitly aimed at migrants or members of the LGBT+ community, while the others are targeted at another *commenter* in the discussion thread. As shown in Figure 5, for both topics and languages, comments targeted at *migrants* or *LGBT+* are generally more emotional than comments targeted at *interlocutors* (fellow-commenters in the discussion thread).

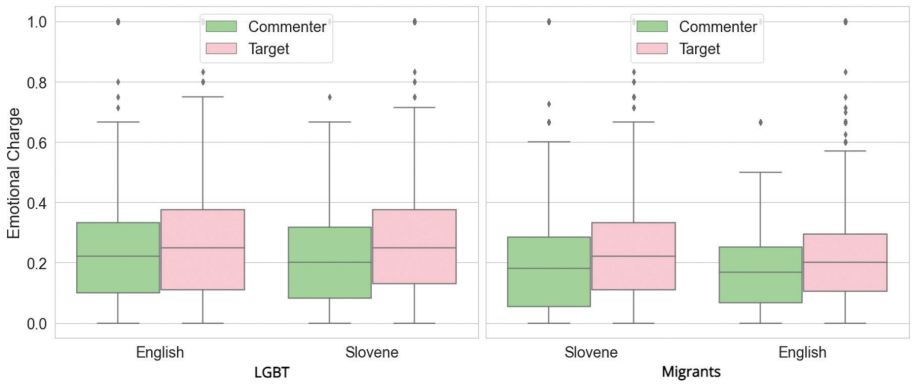

Figure 5: Comparison of emotional charge for different targets of the SUD comment, namely Commenter or Target (LGBT+ persons/migrants), between Slovene and English. The figure shows distributions (rectangles) and variance (lines).

Different topics provoke different emotions. In order to extract the data for a specific emotion, we calculated how much a specific emotion contributes to the total emotional charge. Then, by having the emotional charge distribution of each particular emotion, we were able to compare their manifestation for the two different topics: Migrants and LGBT+. On average, we observed that in English comments users manifest more DISGUST and JOY for the topic of LGBT+, and more SADNESS, FEAR and SURPRISE for Migrants (Figure 6). In Slovene, users manifest significantly more ANTICIPATION and JOY for LGBT+, while for the topic of Migrants, they manifest more ANGER and FEAR (Figure 6). It is interesting to observe that in both languages the LGBT+ topic invokes more

JOY, while the Migrants topic invokes more FEAR. It is also quite evident that emotions are not homogenous for all four subset combinations, with TRUST and FEAR being the most dominant emotions with more than 15% of the total emotional spectrum, while SURPRISE is the least present, taking less than 5% of the emotional spectrum.

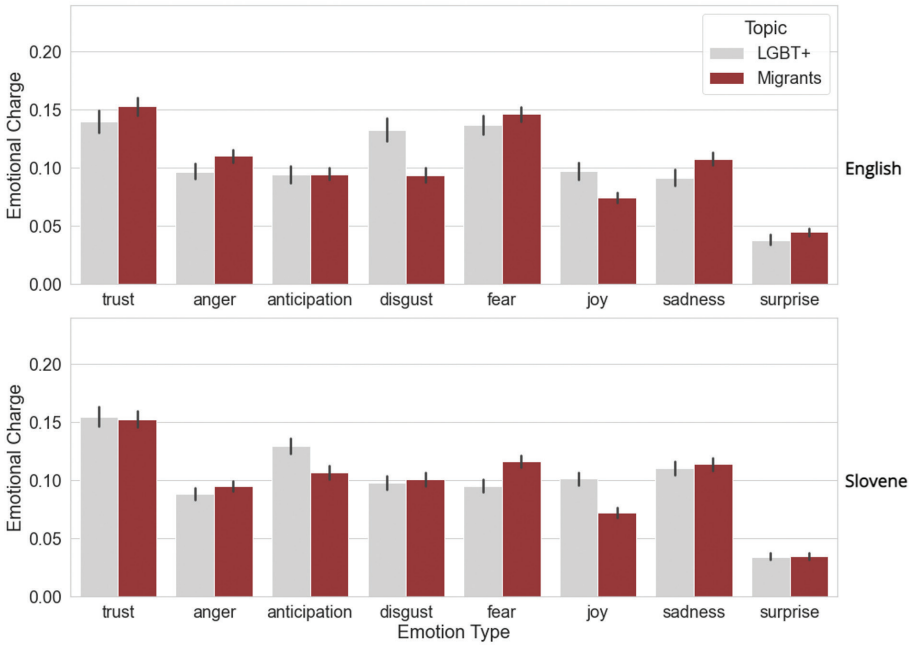

Figure 6: Distribution of emotions in English and Slovene SUD comments for the topics of LGBT+ and Migrants. The figure shows averages (bars) and their confidence intervals (95%) for each emotion present in the NRC Lexicon.

4.2 Emotional Words Analysis

In order to better understand the above quantitative analysis and have a closer look at the investigated data, we performed a qualitative analysis of the emotional words in the corpus.

Table 3 shows the three most frequent emotional words for each language, topic and emotion with the purpose of understanding which words are most commonly connected to which emotions. Some differences and similarities among topics and languages are observed. For the LGBT+ topic, the English commenters seem to be more religion-oriented, frequently using words such as *God*, *disgusting* and *sin*.

Meanwhile, the Slovene take the discussion to a more family-oriented field, using words such as *mother*, *child* and *nurture*. This could be due to the referendum in Slovenia for legalising same-sex marriage that took place in the same period as the data was harvested and has heavily influenced the discussions under Facebook posts by the Slovene media on this topic at the time, where people against framed their arguments around the notion of the traditional family unit, expressing concern with children's rights and same-sex couples' adoptions. Interesting enough, the roles are reversed for the Migrants topic, as now it is the Slovene commenters who are more concerned with religion, using words such as *religion* and *God*, possibly showing fear of a different religion. On the other hand, the English commenters seem to feel more physically threatened by the migrants, using words such as *fight*, *kill* and *idiot*. This could be due to the unprecedented migrant wave through the Balkan route that took place in the same period as the data was harvested and has heavily influenced the discussions under Facebook posts by the Slovene media on this topic at the time as Slovenia has never before experienced anywhere near this rate of the migrant influx, while the topic has been present in the UK political and public debates for many decades.

Individual emotions across languages exhibit mostly similar concepts, yet there are some differences. For example, for the Migrants topic, both English and Slovene commenters use similar words to express emotions, in particular, both groups express DISGUST with the word *terrorist* and show FEAR with *immigrant/fugitive*. On the other hand, English commenters express ANGER for the LGBT+ topic in different terms than the Slovene ones. The English commenters use words such as *disgusting, hate* and *sin,* taking the attitude that being LGBT+ is sinful and repulsive, whereas the Slovene ones show hostility towards the target and, once again, concern regarding children with expressions such as *violence, nurture* and *against*.

Expectedly, in both languages, word usage varies depending on the topic. For example, English commenters express SADNESS with the word *problem* for both the Migrants and the LGBT+ topic, suggesting they perceive both topics as an issue. Yet, Slovene commenters show DISGUST differently for the two topics, exposing what bothers them most

for each. As expected, for the Migrants topic words such as *fugitive, terrorist* and *back* occur frequently, while for the LGBT+ topic *gay, garbage* and *nurture* are recurring.

Table 3: *Top three most frequent emotional words in SUD comments for each emotion, divided per language and topic (Slovene words have their translation after the dash)*

	ENG Migrants	ENG LGBT+	SLO Migrants	SLO LGBT+
ANGER	fight (4.45%) hate (3.12%) money (2.96%)	disgusting (4.22%) hate (3.95%) sin (2.98%)	begunec – fugitive (11.84%) proti – versus (3.73%) terorist – terrorist (2.73%)	proti – against (5.29%) nasilje – violence (2.82%) vzgajati – nurture (2.79%)
ANTICIP.	child (7.7%) good (4.71%) time (4.57%)	God (15.78%), marriage (6.56%) sex (6.25%)	otrok – child (7.16%) vera – religion (6.08%) svet – world (4.60%)	otrok – child (25.28%) zakon – marriage (6.06%) svet – world (3.14%)
DISGUST	hate (4.32%) idiot (3.92%), terrorist (3.27%)	disgusting (4.45%) sick (4.23%) hate (4.16%)	begunec – fugitive (14.31%) nazaj – back (7.20%) terorist – terrorist (3.30%)	peder – gay (5.11%) smeti – garbage (3.52%) vzgajati – nurture (3.31%)
FEAR	problem (4.80%), immigrant (4.56%) war (3.99%)	God (12.24%) disgusting (2.99%) hate (2.80%)	begunec – fugitive (9.75%) vojna – war (4.07%) bog – God (3.00%)	nasilje – violence (2.64%) vzgajati – nurture (2.61%) bog – God (2.30%)
JOY	child (8.72%) good (5.34%) money (3.82%)	God (15.15%) love (8.85%) marriage (6.30%)	otrok – child (9.01%) vera – religion (7.65%) bog – God (4.50%)	otrok – child (28.10%) zakon – marriage (7.00%) mama – mother (3.17%)
SADNESS	problem (6.60%) kill (4.63%) leave (4.20%)	sick (4.36%) hate (4.29%) problem (3.23%)	sam – alone (11.57%) begunec – fugitive (10.76%) brez – without (3.47%)	sam – alone (7.73%) peder – gay (3.88%) mama – mother (3.30%)
SURPRISE	good (8.68%) leave (8.67%) money (6.37%)	good (9.6%) Trump (6.72%) marry (6.23%)	dober – good (8.38%) terorist – terrorist (6.51%) lep – beautiful (5.76%)	dober – good (6.59%) dobro – cool (4.73%) lep – beautiful (4.65%)
TRUST	good (3.16%) show (3.13%) religion (3.11%)	God (12.14%) marriage (5.05%) sex (4.81%)	begunec – fugitive (8.20%) vera – religion (4.27%) svet – council (3.23%)	zakon – marriage (5.60%) pravica – right (4.29%) svet – world (2.90%)

Note. Percentages show the absolute frequency of the word with respect to all the words of the specific emotion. E.g., the word fight covers 4.45% of all anger words for the ENG Migrants dataset.

5 Conclusions

In this paper, we have presented a quantitative analysis of emotions in SUD comments in order to obtain an insight into the emotional footprint of this type of discourse. Applying the NRC Emotion Lexicon, we developed a novel metric named Emotional Charge of the comments to analyse SUD. We implemented this simple, yet effective methodology on the most relevant SUD multilingual dataset which also contains

Slovene data, namely the FRENK dataset, which comprises Facebook comments to posts related to the LGBT+ and the Migrants topic. We showed that SUD comments are more emotional than non-SUD. We also presented how emotions differ depending on the topic. For example, according to the emotion lexicon, the LGBT+ topic invokes more JOY, while the Migrants topic invokes more FEAR. When comparing the emotional charge of SUD comments depending on its target, we observed that comments are more emotional when a user targets the group (LGBT+ or Migrants) compared to a fellow commenter they are having an argument with. Furthermore, we also performed a qualitative analysis of the emotional words, which showed some trends in their usage depending on the topic and language. Slovene commenters to LGBT+ posts are very much concerned with children's wellbeing, while the English ones tend to manifest their opposition and disgust. For the Migrants topic, there is a common tendency in both languages of expressing the same emotion with similar words (e.g., DISGUST – *terrorist*; FEAR – *fugitive/immigrant*).

An original contribution of this study is its demonstration of the methodological potential of the lexical approach for identifying emotions in SUD, which has not been used in the Slovene context yet. The research presented in this paper complements international literature in this domain with the use of richly annotated corpora, emotion lexica and quantitative measures, while also adding a qualitative analysis.

The metric of measuring emotional intensity we have proposed in this paper has proved to be useful and insightful in our research, yet its simplicity could potentially oversimplify the highly complex problem of expressing emotions on social media which transcends linguistic expression and is not only highly context-dependent but is also very culturally nuanced, a common shortcoming of lexicon-based approaches. This is why we propose to experiment with context-aware models and metrics in future work that will better be able to take into account the complexity of this type of communication. We also stress the need for in-depth qualitative sociolinguistic analysis to always complement quantitative and automated approaches that will not only critically evaluate the quantitative approaches of such complex and sensitive phenomena but will also ensure that all relevant aspects of the com-

munication reality are considered before interpreting the results, drawing conclusions and making policy recommendations.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency within the national research project »Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society« (J7-8280, 2017–2019), the Slovenian-Flemish bilateral basic research project »Linguistic landscape of hate speech on social media« (N06-0099, 2019–2023), the national research programme »Slovene Language – Basic, Contrastive, and Applied Studies« (P6-0215) and the national research programme »Digital Humanities: Resources, Tools and Methods« (P6-0436).

References

Alm, C., Roth, D., & Sproat, R. (2005). Emotions from Text: Machine Learning for Text-based Emotion Prediction. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, October 2005, Vancouver, Canada* (pp. 579–586). Association for Computational Linguistics. doi:10.3115/1220575.1220648

Al-Saqqa, S., Abdel-Nabi, H., & Awajan, A. (2018). A survey of textual emotion detection. *8th International Conference on Computer Science and Information Technology (CSIT), July 2018* (pp. 136–142). doi: 10.1109/CSIT.2018.8486405

Aman, S., & Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. In V. Matoušek & P. Mautner (Eds.), *Text, Speech and Dialogue, SD 2007. Lecture Notes in Computer Science (Vol. 4629)* (pp. 196–205). Berlin, Heidelberg: Springer.

Assimakopoulos, S., Baider, F. H., & Millar, S. (2017). *Online Hate Speech in the European Union. A Discourse-Analytic Perspective*. Cham: Springer International Publishing.

Brindle, A. (2016). *The Language of Hate. A Corpus Linguistic Analysis of White Supremacist Language.* London and New York: Routledge.

Canales, L., Daelemans, W., Boldrini, E., & Martinez-Barco, P. (2019). EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media

Text. *IEEE Transactions on Affective Computing*. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8758380

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Daelemans, W., Fišer, D., Franza, J., Kranjčić, D., Lemmens, J., Ljubešić, N., Markov, I., & Popič, D. (2020). *The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene*. Slovenian language resource repository CLARIN.SI. https://www.clarin.si/repository/xmlui/handle/11356/1318

Denecke, K. (2008). Using SentiWordNet for Multilingual Sentiment Analysis. *Proceedings of the 24th International Conference on Data Engineering, 7–12 April 2008, Cancun, Mexico* (pp. 507–512).

Fišer, D., Ljubešić, N., & Erjavec, T. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. *Proceedings of the 1st Workshop on Abusive Language Online, ACL 2017, Vancouver, Canada* (pp. 46–51). Association for Computational Linguistics. doi: 10.18653/v1/W17-3007

Franza, J., & Fišer, D. (2019). The lexical inventory of Slovene socially unacceptable discourse on Facebook. *Proceedings of the 7th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora*, *CMC-Corpora 2019, Cergy-Pontoise, France*. Retrieved from https://hal.archives-ouvertes.fr/hal-02292616/document#page=50

Ghazi, D. (2016). *Identifying Expressions of Emotions and Their Stimuli in Text*. PhD dissertation. Canada: University of Ottawa.

Gitari, N. D., Zuping, Z., Hanyurwimfura, D., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering (Vol. 10, No.4)* (pp. 215–230).

Knoblock, N. (2017). Xenophobic Trumpeters: A corpus-assisted discourse study of Donald Trump's Facebook conversations. In A. Musolff (Ed.), *Journal of Language Aggression and Conflict (Vol. 5, No.7)* (pp. 295–322). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ljubešić, N. (2019). *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Slovenian*. Ljubljana: Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1251

Ljubešić, N. (2020). *The CLASSLA-StanfordNLP model for lemmatisation of standard Slovenian 1.1*, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1286

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). *The FRENK datasets of Socially Unacceptable Discourse in Slovene and English*. International Conference on

Text, Speech, and Dialogue. Springer, Cham. doi: 10.1007/978-3-030-27947-9_9

Ljubešić, N., Fišer, D., Erjavec, T., & Šulc, A. (2021). *Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.1*. Ljubljana: Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1462

Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021). Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection. *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 149–159). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.wassa-1.16/

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate Speech Classification in Social Media Using Emotional Analysis. *7th Brazilian Conference on Intelligent Systems (BRACIS), 22–25 October 2018, Sao Paulo, Brazil* (pp. 61–66). doi: 10.1109/BRACIS.2018.00019

Mohammad, S., & Yang T. (2011). Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)* (pp. 70–79). Portland, Oregon: Association for Computational Linguistics.

Mohammad, S., & Turney, P. D. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, Los Angeles, California* (pp. 26–34).

Pahor de Maiti, K., Fišer, D., & Ljubešić, N. (2019). How haters write: analysis of nonstandard language in online hate speech. *Proceedings of the 7th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora*, *CMC-Corpora, 9–10 September 2019, Cergy-Pontoise, France*. Retrieved from https://hal.archives-ouvertes.fr/hal-02292616/document#page=44

Peng Q., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. Retrieved from https://arxiv.org/abs/2003.07082

Plutchik, R. (1980). *Emotion: Theory, research and experience,* 1. Academic Press.

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.

Pratt, J. W., & Gibbons, J. D. (1981). Kolmogorov-Smirnov two-sample tests. *Concepts of nonparametric theory*. Springer, New York, NY. 318–344.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. doi: 10.1037/h0077714

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 695–729. doi: 10.1177/05390184050582

Vehovar, V., Povž, B., Fišer, D., Ljubešić, N., Šulc, A., & Jontes, D. (2020). Družbeno nesprejemljivi diskurz na Facebookovih straneh novičarskih portalov. *Teorija in Praksa*, *57*(2), 622–645.

Zad, S., Jimenez, J., & Finlayson, M. A. (2021). Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon. *Proceedings of the 5th Workshop on Online Abuse and Harms, 6 August 2021, Bangkok, Thailand* (pp. 102–111). Retrieved from https://aclanthology.org/2021.woah-1.pdf

Analiza čustev v družbeno nesprejemljivem diskurzu

Besedila pogosto izražajo avtorjevo čustveno stanje in pokazalo se je, da imajo informacije o čustvih potencial za odkrivanje in analizo sovražnega govora. V prispevku predstavljamo kvantitativno metodologijo analize čustev v besedilu. Na podlagi leksikona čustev NRC Emotion Lexicon in Plutchikovega modela osmih osnovnih čustev smo definirali preprosto, a učinkovito metodo za odkrivanje čustvene zaznamovanosti besedila. Z navedeno metodologijo smo raziskali čustveno zaznamovanost besedil, označenih kot družbeno nesprejemljivi diskurz (DND), ki predstavlja izrazito in potencialno škodljivo vrsto besedila ter se dandanes hitro širi na družbenih omrežjih. Metodo čustvene zaznamovanosti smo aplicirali na korpus komentarjev s Facebooka. Primerjavo in analizo smo izvajali na štirih zbirkah podatkov v dveh jezikih, in sicer v angleščini in slovenščini, ter na dveh temah, pravice LGBT+ skupnosti in evropska migrantska kriza. Ugotovili smo, da je vsebina DND komentarjev bistveno bolj čustvena od tistih, ki ne vsebujejo DND. Poleg tega smo pokazali razlike v izražanju čustev glede na jezik, temo in tarčo komentarjev. Izsledke kvantitativne metodologije analize čustev smo podprli s kvalitativno analizo korpusa, kjer smo preučili najpogostejše čustveno zaznamovane besede, povezane z vsakim čustvom v vseh štirih zbirkah podatkov. Ugotovili smo, da se čustveno zaznamovane besede v DND bistveno razlikujejo glede na temo, medtem ko obstaja med jeziki precejšnje prekrivanje.

Ključne besede: čustva, družbeno nesprejemljivi diskurz (DND), sovražni govor, družbena omrežja, korpusi