The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene

Nikola Ljubešić
Dept. of Language Technologies
Jožef Stefan Institute, Slovenia
nikola.ljubesic@ijs.si

Ilia Markov
CLIPS Research Center
University of Antwerp, Belgium
ilia.markov@uantwerpen.be

Darja Fišer
Faculty of Arts
University of Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Walter Daelemans
CLIPS Research Center
University of Antwerp, Belgium
walter.daelemans@uantwerpen.be

Abstract

In this paper, we present emotion lexicons of Croatian, Dutch and Slovene, based on manually corrected automatic translations of the English NRC Emotion lexicon. We evaluate the impact of the translation changes by measuring the change in supervised classification results of socially unacceptable utterances when lexicon information is used for feature construction. We further showcase the usage of the lexicons by calculating the difference in emotion distributions in texts containing and not containing socially unacceptable discourse, comparing them across four languages (English, Croatian, Dutch, Slovene) and two topics (migrants and LGBT). We show significant and consistent improvements in automatic classification across all languages and topics, as well as consistent (and expected) emotion distributions across all languages and topics, proving for the manually corrected lexicons to be a useful addition to the severely lacking area of emotion lexicons, the crucial resource for emotive analysis of text.

1 Introduction

Emotion lexicons are rather scarce resources for most languages (Buechel et al., 2020), although they are a very important ingredient for robust emotion detection in text (Mohammad et al., 2018). They are mostly differentiated between depending on the way they encode emotions – either via continuous or discrete representations (Calvo and Mac Kim, 2013). In this work, we present emotion lexicons for three languages – Croatian, Dutch and Slovene, developed by manually correcting automatic translations that are part of the NRC Emotion Lexicon (Mohammad and Turney, 2013). In that lexicon, sentiment (positive or negative) and a discrete model of emotion covering *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*, are encoded via a binary variable for each emotion. The size of the lexicon is 14,182 entries, and automatic translations of that lexicon were performed by the authors in 2017 via Google Translate.

The three languages in question are generally not covered well with emotional lexicons. The first lexical dataset encoding emotion for Croatian was published recently (Ćoso et al., 2019), encoding valence and arousal for 3,022 words. For Dutch two resources exist – the LIWC (Tausczik and Pennebaker, 2010) translation into Dutch (Boot et al., 2017) covering around 4,500 entries, and the norms for 4,300 Dutch words for dominance, valence and arousal (Moors et al., 2013). There are no available emotion lexicons for the Slovene language except the automatic translation of the NRC lexicon.

Just recently, an automatic approach to building emotion lexicons for many languages beyond mere translation has emerged (Buechel et al., 2020) and its results cover all three languages in question, showing reasonable correlations with the available small psychological norms in various languages.

2 The Lexicon Translation Procedure

We base our approach to building reliable emotion lexicons for the three languages in question on a very simple approach – we manually correct automatic translations of an existing English emotion lexicon.

Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media, pages 153–157
Barcelona, Spain (Online), December 13, 2020.

These automatic translations consist to the greatest part of single-word translations, only rare entries being translated with short phrases.

We used different translators for each language as our aim was for the translations to be performed by native speakers. We used a uniform translation methodology across the three languages, ensuring comparability between the resulting lexicons. The guidelines presented to the translators were the following:

- While translating, the sentiment and emotion labels should be taken into account by the translator.
- If a source word is polysemous, we translate only those senses that relate to the given sentiment and emotion.
- We include all target synonyms of the source word that are frequent and that are not polysemous or homonymous with target words of a different sentiment and emotion association.[1]
- Given that the original lexicon was not part-of-speech encoded, we do not encode it in the translations either. We add translations in multiple parts-of-speech if this follows the guidelines above.

While we decided to correct translations of all entries for Croatian and Slovene, for Dutch we only corrected those entries that were marked for emotion, i.e., they had at least one positive value among the two sentiment and eight binary variables encoded in the lexicon. The reason for this decision was the limited availability of the Dutch translator and the fact that the emotion-marked part of the lexicon is most useful for various applications. Given this decision, the resulting lexicons of Croatian and Slovene have 14,182 entries, while the Dutch version has only 6,468 entries.

Translating the lexicons resulted in a different level of modifications in each of the three languages. In Table 1, we present the distribution of emotion-marked (left) and emotion-empty (right) entries regarding whether (1) the original translation was kept, (2) the original was extended with additional translations, or (3) the entry was fully modified. Among the marked entries the fewest interventions were necessary in the Dutch lexicon. This might be due to two facts: (1) the English-Dutch automatic translation is performing in general better than the English-Croatian and English-Slovene one, and (2) the Dutch inflectional system is closer to the English one than that of Croatian and Slovene, the latter both being (South) Slavic languages. For the emotion-empty entries that were translated to Croatian and Slovene only, these have required overall less intervention when compared to emotion-full entries. A higher number of emotion-empty entries was kept in their original state and a lower number of these entries was fully modified. This phenomenon might derive from the fact that emotion-marked words are to some level culture-specific (Jackson et al., 2019; Markov et al., 2018), but even more to our intuition that emotion-marked words are harder to machine-translate than those that are emotion-empty.

	marked			empty		
	original	extended	fully modified	original	extended	fully modified
Croatian	45.9%	22.7%	31.5%	54.8%	16.5%	28.7%
Dutch	68.2%	19.1%	12.7%	-	-	-
Slovene	31.0%	7.8%	61.2%	35.9%	7.2%	56.9%

Table 1: The per-language distribution of emotion-loaded (left) and empty (right) entries that were kept untouched, were only extended, or were fully modified during the manual correction process.

3 Analyses

In this section, we present two types of analyses of the manually corrected LiLaH emotion lexicons – one based on machine learning that compares the original automatically translated NRC and the manually corrected LiLaH lexicon, and another based on descriptive statistics, exploiting only the LiLaH emotion lexicon.

[1]The percentage of entries translated into more than a single translation is the following by language: Croatian has 27% of such translations, Dutch 24%, and Slovene 24%.

In the first analysis, we measure the impact of the corrections performed in the lexicon on the task of supervised classification of messages potentially containing socially unacceptable discourse (SUD; hate speech being one type of SUD).

In the second analysis, we perform a descriptive analysis of emotion difference between socially unacceptable and socially acceptable discourse (SUD and non-SUD onwards), showcasing the potential of the new lexicon for data analysis.

We perform both analyses on the same set of datasets with manual annotations of SUD - the FRENK Croatian, English and Slovene datasets and the LiLaH Dutch dataset. All four datasets consist of comment threads of mainstream media Facebook posts in Croatia, Great Britain, Slovenia and Flanders. The comment threads in the datasets belong to two topics: migrants and LGBT. The size of the datasets varies, per language and topic, between 3836 and 6941 comments. Each comment in a thread is annotated with an annotation schema encoding the type and target of SUD (Ljubešić et al., 2019). In this work, we only use the basic information whether a comment contains SUD or not. The percentage of SUD comments in the datasets lies between 32% (Dutch LGBT) and 64% (Croatian LGBT).

3.1 Machine Learning Experiments

For our machine learning experiments we transform each text in the datasets in a frequency distribution of emotion associations (or their lack thereof). Let us assume that we discriminate only between two emotions, *anger* and *joy*. For a textual instance "I am so happy", which we lemmatize into a sequence "I be so happy", in our lexicon we cover only the word "happy" with *joy* associated, but not *anger*. We represent this utterance, given that there are four words in the utterance, as {`anger_yes:0,anger_no:4,joy_yes:1,joy_no:3`}.[2]

We perform our experiments by training a binary SVM classifier with an RBF kernel over the textual data transformed with (1) the automatic NRC lexicon and (2) the manually corrected LiLaH lexicon. In Table 2, we present the results of the classification experiments over the three languages of the LiLaH lexicon, and the two topics, migrants and LGBT. We evaluate each system via macro F1 by cross-validating. As a lower bound we use the stratified random baseline. We finally calculate the Δ as the percentage of error reduction that was obtained by manually correcting the machine translated lexicon, considering the random baseline results as our lower bound. All bold results are statistically significantly higher than their counterpart given the McNemar test (McNemar, 1947) at a significance level of $p < 0.0001$.

topic	lang	random	automatic	manual	Δ
migrants	sl	0.495	0.614	**0.639**	4.9%
lgbt	sl	0.500	0.581	**0.636**	11.0%
migrants	nl	0.501	0.601	**0.646**	9.0%
lgbt	nl	0.503	0.617	**0.648**	6.2%
migrants	hr	0.501	0.625	**0.648**	4.6%
lgbt	hr	0.498	0.619	**0.636**	3.4%

Table 2: The macro F1 results of the binary supervised classification experiments on using the automatically translated lexicon or the manually corrected one.

The results show that the manually corrected lexicons improve over the system using automatic translation in each of the six (three languages, two topics) cases, all the improvements being statistically significant. The percentage of the error reduction obtained by moving from the automatically translated lexicon to the manually corrected one, using the random baseline as a lower bound, lies between 3.4% and 11%. These results show that the manually corrected lexicons contain more signal than the automatically translated ones for the task at hand. One has keep in mind that machine learning was used here not to obtain best possible results on the task, but just to measure the impact of the improvements of the lexicon on the formal description of the task at hand, and thereby on other similar problems.

[2]Different representations of the data, e.g., recording only positive associations, not negative ones, resulted in slightly lower results, but yielding the same conclusions.

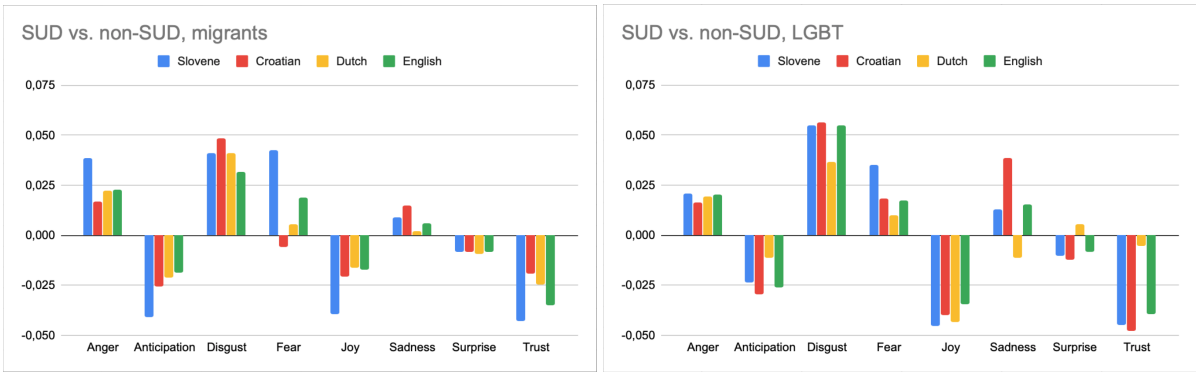

Figure 1: Difference in emotion distribution between texts containing socially unacceptable discourse and not containing such discourse, across the two topics (migrants left, LGBT right) throughout the four languages.

3.2 Emotion Difference Analysis

In this subsection, we do not focus on the difference between the automatically translated and the manually corrected lexicon, but simply apply the manually corrected lexicon over the FRENK and LiLaH datasets to showcase both its usefulness and robustness. Given that in this analysis we are not interested in the difference between the automatically and manually translated lexicon, we are able to include English into this analysis and perform it on all four languages.

We measure the emotion difference between the text containing and not containing SUD as the difference in their probability distribution of emotion associations. For each of the two categories we count the number or tokens that are associated with a specific emotion, and then calculate the probability distribution over the eight emotions. Finally, we simply subtract the distribution calculated on texts not containing SUD from the distribution calculated on texts containing SUD. By doing so we discard the problem of some emotions being more represented in the underlying lexicons than others. The resulting emotion differences are presented in Figure 1 separately for the topic of migrants and LGBT throughout the four languages.

The results show in the first place a high consistency between the four languages. Overall, on both topics, *anger*, *disgust*, *fear* and *sadness* are more prevalent in SUD texts, while the remaining four emotions are more prevalent in non-SUD texts.

The most prominent differences between the four languages can be observed on the emotion of *fear* for the topic of migrants, with Croatian expressing an even higher level of *fear* among the non-SUD texts, which is probably due to a very hardlined discussion of the topic. Another notable difference on the series of emotions can be observed on the topic of LGBT for the Dutch language, which contains an overall less aggressive discussion on that topic in comparison to the three other languages.

Regarding the differences between the two topics, SUD texts on the LGBT topic seem to contain more *disgust* and *sadness* and less *joy*, while SUD texts on the topic of migrants are imbued with more *anger*.

4 Conclusion

In this paper, we presented the LiLaH emotion lexicons – manually corrected automatic translations of the NRC Emotion Lexicon into Croatian, Dutch and Slovene. We have shown that the manually corrected lexicons are more potent in a supervised learning scenario where the lexicon information is used for data representation. We have additionally showcased a usage scenario of the LiLaH lexicons for emotion text analysis – we analysed emotional differences between texts containing socially acceptable and unacceptable discourse, showing a large amount of consistency across languages and topics, with some interesting peculiarities. The presented lexicons are available for download at `http://hdl.handle.net/11356/1318` (Daelemans et al., 2020).

Acknowledgement

This work has been supported by the Slovenian Research Agency and the Flemish Research Foundation through the bilateral research project ARRS N6-0099 and FWO G070619N "The linguistic landscape of hate speech on social media", the Slovenian Research Agency research core funding No. P6-0411 "Language resources and technologies for Slovene language", and the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (grant no. 875263).

References

Peter Boot, Hanna Zijlstra, and Rinie Geenen. 2017. The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1):65–76.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and Evaluating Emotion Lexicons for 91 Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online, July. Association for Computational Linguistics.

Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Bojana Ćoso, Marc Guasch, Pilar Ferré, and José Antonio Hinojosa. 2019. Affective and concreteness norms for 3,022 Croatian words. *Quarterly Journal of Experimental Psychology*, 72(9):2302–2312.

Walter Daelemans, Darja Fišer, Jasmin Franza, Denis Kranjčić, Jens Lemmens, Nikola Ljubešić, Ilia Markov, and Damjan Popič. 2020. The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene. Slovenian language resource repository CLARIN.SI.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In *International Conference on Text, Speech, and Dialogue*, pages 103–114. Springer.

Ilia Markov, Vivi Nastase, Carlo Strapparava, and Grigori Sidorov. 2018. The role of emotions in native language identification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 123–129, Brussels, Belgium. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior research methods*, 45(1):169–177.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.