Social-Emo.Sr: Emotional Multi-Label Categorization of Conversational Messages from Social Networks X and Reddit

Milena Šošić, Ranka Stanković, Jelena Graovac

mr Milena Šošić
doktorand na Matematičkom fakultetu Univerziteta u Beogradu
E-mail: milena.sosic@gmail.com, pd202030@alas.matf.bg.ac.rs

prof. dr Ranka Stanković
vanredni profesor na Rudarsko-Geološkom fakultetu Univerziteta u Beogradu
E-mail: ranka.stankovic@rgf.bg.ac.rs

prof. dr Jelena Graovac
vanredni profesor na Matematičkom fakultetu Univerziteta u Beogradu
E-mail: jgraovac@matf.bg.ac.rs

Social-Emo.Sr: Emotional Multi-Label Categorization of Conversational Messages from Social Networks X and Reddit

In the digital environment of South Slavic languages, emotion analysis in texts on social media is becoming increasingly important for understanding public opinion, creating personalized content, and analyzing user interactions. This presentation presents a detailed methodology and results of corpus annotation in the Serbian language according to Plutchik's categorization model, which identifies eight basic emotional categories: joy, sadness, anger, fear, trust, disgust, anticipation, and surprise. The aim of the research is to analyze the emotional content of texts taken from social media X (formerly Twitter) and Reddit, each collection containing around 17,000 individual messages and approximately 5,000 complete conversations. The corpus annotation process involved several stages: data collection and preparation, manual annotation by experts, verification of annotation accuracy, and statistical analysis of the harmonized labels. By using a multi-label annotation approach, a richer and more qualitative analysis of emotional states was made possible, with particular significance for the application in analyzing complex emotional content found on social media.

To collect data, automated tools were used to download conversations written in Serbian from social media accounts that address current social, political, musical, and sports topics. Data preparation involved additional selection of messages to ensure the quality of their content, while maintaining the conversational structure of the retrieved data. During data preparation, messages were preliminarily annotated using automatic methods, employing both classical and advanced computational linguistics techniques to improve the efficiency of the manual labeling process. Teams of linguists and psychologists reviewed and assessed the automatically assigned labels for their validity concerning the textual content to which they were assigned. To ensure high accuracy and consistency, standardized procedures were used for training annotators and verifying their evaluations through statistical measures of annotation reliability. The analysis of annotation reliability demonstrated that it is possible to classify emotions in texts from social media in Serbian using Plutchik's model. Statistical data analysis revealed significant distributions of emotions in the messages and provided insights into users' emotional reactions to various emotional stimuli and thematic content.

The multi-label categorized emotional corpus in Serbian Social-Emo.SR represents a significant advancement toward a deeper understanding of emotional dynamics on social media among users. In addition to enriching linguistic resources for the Serbian language, this corpus opens new possibilities for application in research, commercial applications, and enhancing mental health analysis of the population. The potential application of modern methodologies on the developed corpus would enable the creation of useful tools for recognizing and reflecting

the complexity of human emotions in the current digital world within the Serbian-speaking community. The corpus will be published under open license CC-BY-4.0.

Keywords: *emotions, Plutchik's model, annotation, corpus, social media, Serbian language*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Mihailo Škorić
Društvo za jezičke resurse i tehnologije JeRTeh
E-mail: mihailo@jerteh.rs

New Language Models for South Slavic Languages

The report will present the challenges and perspectives of modeling South Slavic languages, especially the general language models built on the transformer architecture (BERT, GPT), available corpora of texts for training those models, and the quantity and quality of those corpora. The presentation will offer an overview of the available data and models, primarily the latest textual corpora. The first corpus, *Umbrella*, represents the umbrella web corpus of South Slavic languages and at the same time the largest corpus of texts in the region, includes all other currently available regional web corpora and contains over eighteen billion words. The second corpus, *S.T.A.R.S*, gathers academic works written in the Serbian language and includes, most notably, eleven thousand dissertations downloaded from the NARDUS platform, and a large number of scientific and professional works downloaded from various open repositories that are included in the *eScience* system. In addition, academic corpora of other South Slavic languages will be discussed, which were created from works stored on various web platforms: DABAR (for the Croatian language), the repositories of the universities in Maribor, Ljubljana, Primorska and Nova Gorica, and the DiRROS and REVIS repositories (for the Slovene language), the repository of the universities in Zenica, Sarajevo and East Sarajevo (for the Bosnian language), the repository of the University of Goce Delčev and St. Kliment Ohridski (for the Macedonian language) and the repository of the University of Montenegro (for Montenegrin). Finally, we will talk about new models for text vectorization in South Slavic languages, which were trained using the aforementioned corpora. An analysis of their performance on a number of previously established tasks will be presented, with reference to the model performance and improvements over models trained on the previous generation of the corpora.

Keywords: *Large text corpora, language models, South Slavic languages*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

811.163'322(048)(0.034.2)
004.8(048)(0.034.2)

INTERNATIONAL Conference South Slavic Languages in the Digital Environment JuDig (2024 ; Beograd)

Book of abstracts [Elektronski izvor] / International Conference South Slavic Languages in the Digital Environment JuDig, International Conference South Slavic Languages in the Digital Environment JuDig, November 21-23. 2024, [Belgrade] ; [organisers University of Belgrade - Faculty of Philology [and] Society for Language Resources and Technologies JeRTeh]. - Belgrade : University, Faculty of Philology, 2024 (Belgrade : University, Faculty of Philology). - 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovnog ekrana. - Tiraž 100.

ISBN 978-86-6153-754-7

a) Јужнословенски језици -- Рачунарска лингвистика – Апстракти

COBISS.SR-ID 157072905