TWITTER BEFORE X: SCENES FROM THE BALKANS

Filip Dobranić,[1] Nikola Ljubešić[1, 2]
[1]Institute for Contemporary History
[2]Department of Knowledge Technologies, Jožef Stefan Institute

We present a corpus of over 170 million Twitter posts in Slovenian, Croatian, Bosnian, Serbian and Montenegrin, collected between 2017 and 2023. After describing the data collection, our focus moves to the challenges of the division and linguistic processing of the data given that we deal with an area as ethnically and linguistically interconnected as the Balkans. An exploratory analysis into the quality, quantity and differences between the collections provides evidence of real-life events' effect on online user-generated content production, most no- tably measures mitigating the spread of COVID-19 and elections. We investigate the use of emotionally charged words through time and provide a corpora-wide overview of the most prolific authors, hashtags, and mentions before conclud- ing the paper with an invitation for further research into this vast and diverse contemporary-historical corpus of online speech.

Keywords: Twitter corpus, Slovenian, HBS macro-language, distant reading

1 INTRODUCTION

Social media platforms have become an integral part of modern communication, providing a vast source of data for researchers to study language use and societal trends. While collections of tweets have been used for various tasks, including forecasting mental illness among users (Reece et al., 2017), the majority of research focuses on Twitter's interplay with (inter)national politics (Karami et al., 2020). The wide range of uses of Twitter corpora highlight not only their importance for understanding contemporary politics ranging from political campaigns by institutional actors to the organisation of grass roots protests (Altman, 2018; Bronstein et al., 2018; Chan & Yi, 2024; Effing et al., 2011; Flew & Iosifidis, 2020;) but also their value when it comes to understanding other societal structures such as traditional media (Zhao et al., 2011) and studying human affect and social media behaviour more broadly (Wang et al., 2020; Y. Qi & Shabrina, 2023).

In this paper, we present a collection of corpora of tweets written in Slovenian and any of the languages of the ISO-639-3 HBS macro-language (the macro-language including Croatian, Bosnian, Montenegrin and Serbian) tweets collected between the years 2017 and 2023. The corpora are the result of processing $172,806,656$ tweets produced by $2,959,373$ authors, containing $3,418,701$ hashtags, and $5,405,589$ mentions.

The paper is structured as follows. In the following section we outline the process of data collection, filtering based on language (2.1), redaction (2.2), and linguistic annotation (2.3) of the tweets. We follow up by discussing the split of our collection into three corpora (2.4) and outline some of the optimisations to speed up processing (2.5). Next, we present the corpora structure in Section 3, providing a short distant reading and analysis of the corpora before concluding with an invitation for future research and use of the corpora.

2 CORPUS CREATION

The data collection took place between the years 2017 and 2023. The tweets were collected with the TweetCaT methodology as outlined in (Ljubešić et al., 2014) using seed terms to identify users tweeting in the desired language, then storing all the tweets by identified users in JSON files. Two collection processes took place simultaneously: one with Slovenian seed words and one with seed words in HBS languages. After the collection period came to a close, the tweets were filtered through a secondary sieve to exclude tweets and retweets in languages we were not interested in. While a user might predominantly tweet in their preferred (often native) language, it is quite common for people to share content and converse with other users on the internet in various languages.

2.1 Language filtering

After collecting the tweets based on our prediction of users' native language, we had to filter out the tweets and retweets that contained either no recognisable language or were predominantly in a language we were not interested in. For the Slovenian corpus, we relied on Twitter's own language identification. The

API responses contained a key with the predicted language, and we used that to discard any tweet not marked as Slovenian by Twitter.

For HBS languages we found Twitter's labelling lacking. While there were few false positives, our initial investigation found a significant amount of tweets written in HBS languages that were mislabelled or labelled as unknown. In order to address the shortcomings of Twitter's language identification, we used FastText (Bojanowski et al., 2017) to determine the tweets' language. Our goal was to find tweets in any of the HBS languages. We hypothesised that if FastText gives multiple low-confidence predictions of HBS languages we can treat that as a case of a relatively high confidence prediction that it is written in at least one of the languages.

In order to confirm our hypothesis we annotated a random sample of 500 tweets and tested FastText's language identification, changing two defined hyper-parameters of the classifier (number of guesses and confidence threshold). The tests confirmed the viability of our approach (multiple low-confidence predictions were a good signal for the tweet being in one of the HBS languages), but showed that the recall and accuracy of our approach depends on hyper-parameters used with FastText: the number of guesses and required confidence threshold.

Consequently, we used our set of 500 annotated tweets to tune the two hyper-parameters for the language classifier. We then investigated the results in bands based on their performance characteristics. You can consult our code at https://dihur.si/muki/twitter for specifics. After investigating hyper-parameter configurations based on their precision and recall on the annotated dataset, we opted for the hyper-parameter configuration ensuring the highest recall, since a relatively marginal maximum gain in precision would mean 30% fewer tweets in the corpus. We prioritised more tweets in our corpora even if they contain a slightly larger amount of tweets not in the target languages as opposed to a more narrow corpus with higher language guarantees. For further research requiring stricter limits on language presence, researchers can always further filter the current tweets to a degree appropriate for their research.

2.2 Redaction

Twitter's API responses present a lot of structured information we are not necessarily interested in, e.g. coordinates, favorite count at time of retrieval, sensitive content flag, extracted entities, user metadata etc. We opted to redact this information both out of privacy concerns as well as a mindfulness of the corpus' final size and resources required to process it.

While redacting, we used the opportunity to unify the structure of the data. Since the collection took place on such a large time scale, the API responses themselves changed over time. The most notable of these is the key containing the tweet's text, which changed from *full_text* to *text*, along with at least on other relevant property, *truncated*, denoting responses that did not contain the full tweet text. In our redacted data structure we assign the tweet's text to the key *text* and retain the *truncated* boolean flag. The structure of our redacted tweet data follows:

- *created_at* ($string$) retains the timestamp of tweet creation provided by the API
- *truncated* ($boolean$) marks truncated tweets with *true*
- *user_screen_name* ($string$)
- *text* ($string$)
- *id_str* ($string$) retains the tweet's ID number as string
- *is_retweet* ($boolean$) marks tweets that are retweets with *true*
- *source_tweet_id_str* ($string|null$) if the tweet was a retweet, this contains the id of the original tweet, otherwise null

2.3 Linguistic annotations

Once the tweets were redacted, we proceeded to linguistically annotate them. We performed the annotation automatically with the CLASSLA-Stanza pipeline (Terčon & Ljubešić, 2023), a fork of the Stanford Stanza pipeline (P. Qi et al., 2020). We prefer the CLASSLA pipeline over Stanza since the former's models are based on a larger training dataset, use large inflectional lexicons, support both standard and Internet-non-standard language, and have support for Named Entity Recognition (NER). Since language captured on the internet often differs from the formal version (e.g. there is no *@mention* in Slovenian), we used

the non-standard pipelines by applying the *type="nonstandard"* argument in the pipeline.

2.4 Language considerations and the splitting of the HBS corpus

A regular stumbling block when dealing with data in the HBS macro-language, especially when the data were collected in the wild, either on the web or from social media, is whether, and if so, how to further divide content written in the underlying languages. Given the shortness of messages on Twitter, it is a proper challenge to perform a reasonable automated job on this task (Ljubešić & Kranjčić, 2014). Most proposed solutions are based on machine learning on available data, which are known to overfit to the training data, giving perfect results if the test data are similar, but far from perfect results otherwise (Rupnik et al., 2023).

The CLASSLA-Stanza pipeline does not have an HBS pipeline, but either a Croatian or a Serbian pipeline. The biggest difference between these two pipelines, or rather, the data these pipelines were trained on, are that the Serbian pipeline was trained on ekavian data (*lepo*, *beži*), while the Croatian pipeline was trained on standard and non-standard ijekavian data (*lijepo*, *bježi*). Furthermore, the Croatian non-standard data cover most phenomena specific to the Bosnian and Montenegrin language, such as the synthetic future tense (*smetaću* vs. *smetat ću* in standard Croatian), or the usage of both *što* and *šta* pronouns (standard Croatian allows only the former), etc. The Croatian non-standard processing pipeline is capable of dealing with most of the lexical differences between Croatian, Bosnian and Montenegrin (in Croatian verbs ending in *-irati* mostly end in the other langauges in *-isati* and *-ovati* etc.) due to static embeddings used in the pipeline that have been trained on web data, where most variants can be found.

Important to note is also that the Croatian pipeline lemmatizes into the ijekavian variant, while the Serbian pipeline lemmatizes into the ekavian variant.

Given the above described situation, we have decided not to continue discriminating between the four languages contained in the HBS macro-language as such a division would result in significant errors, but rather to be application-oriented and follow the division by the most prevalent linguistic difference

between the different standards in the macro-language, which is also reflected in the CLASSLA-Stanza processing pipeline, namely the ekavian vs. ijekavian variant. For that reason we have used a lexicon-based approach from our previous work (Ljubešić et al., 2018) to classify users either as using the ekavian or the ijekavian variant of the HBS macro-language, constructing two separate corpora called HBS-ekavian and HBS-ijekavian. Each corpus was processed linguistically with the corresponding pipeline, the HBS-ekavian corpus with the Serbian pipeline, and the HBS-ijekavian corpus with the Croatian pipeline.

Since the CLASSLA-Stanza pipeline is capable of processing the Latin script only for both the HBS-ekavian and HBS-ijekavian corpus, we used the *cyrtranslit* (Labrèche, 2023) transliteration tool to transform tweets from the Cyrillic into the Latin script. We retained the original text for future research endeavours and annotated all the tweets in our collection with a boolean value denoting whether or not the tweet was transliterated. Tweets that had more than 20% of their characters modified after transliteration had the transliterated flag set to true (the rest have the flag set to false) to allow for easier filtering of the most heavily transformed tweets.

2.5 Optimisations for linguistic annotations with CLASSLA-Stanza

While individual tweets technically represent individual “documents” in our corpora, due to CLASSLA’s startup time and the tweets’ relatively short length, processing each and every one of them individually would take a prohibitively long time. To mitigate this, we collect all the tweets published in a single day, join their texts with ”\n|\n”, process the day’s worth of tweets as one document, and then split the result at the paragraph containing the pipe character ”|”. Our ad hoc benchmarks showed a reduction in processing time by a factor of 8.

Since a significant part of the corpora are retweets, we performed another optimisation during our processing. For every day, we begin by only processing original content first, storing retweet IDs along the way. Once we iterate through all the day’s tweets and process the originals, we process all unique retweeted tweets, then copy the results to their corresponding retweets. This ensures that we processed a tweet at most twice in a day instead of every time we encounter a retweet.

3 CORPUS OVERVIEW AND ANALYSIS

Finishing the steps outlined above, we ended up with our three linguistically annotated corpora: the Slovenian corpus, the HBS-ijekavian corpus, and the HBS-ekavian corpus. In this section we outline the structure of the corpus and perform some preliminary analysis to showcase the corpus' ability to aid in answering a wide and diverse range of research questions. We begin with an analysis of the corpora as a whole, then investigate specific authors, hashtags, and mentions. Basic size metrics are presented in Table 1.

Table 1: Metadata on each of the three corpora.

What	*Slovenian*	*HBS-ijekavian*	*HBS-ekavian*
Total number of tweets	42,483,342	31,199,242	99,124,072
Original tweets	23,293,074	26,975,269	64,986,741
Retweets	19,190,268	4,223,973	34,137,331
Truncated tweets	1,137,603	1,006,906	3,123,005
Authors	483,216	601,282	1,874,875
Hashtags	1,012,474	963,748	1,442,479
Mentions	736,573	1,527,489	3,141,527

3.1 Size and activity

In this section we discuss the size and tweet production through time for the three corpora. While Twitter usage fluctuates on a monthly basis, there are a few notable drops or rises in production visible from this perspective. We discuss these for each of the corpora.

A total of $42,483,342$ tweets were processed to form the Slovenian corpus, of those $23,293,074$ (almost 55%) represent original content, the rest are retweets. It's important to note that this includes only content written in the target language (Slovenian). If a user retweeted content in other languages the posts are ignored. The number of tweets and retweets through time can serve as an approximation of Twitter's relative speed of growth (or contraction), but it does not accurately estimate the absolute quantity of posts produced by Twitter users.

Looking at the data in terms of total Slovenian tweet production, we observe an order of magnitude increase of number of tweets (both original and retwets) in

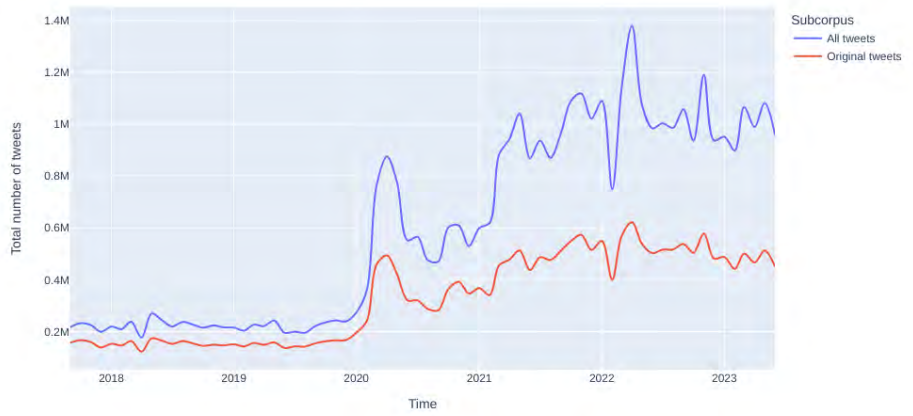

Figure 1: Number of monthly tweets in the Slovenian corpus through time.

March 2020, coinciding with the official beginning of the COVID-19 pandemic in Slovenia and first "lockdowns". During the summer months of the same year, as quarantine measures were loosened and people moved back to meatspace, we can observe a dip in posting activity. While remaining relatively low until March 2021 total number of tweets still remains above levels before the COVID-19 pandemic. In March and April 2021 we observe another rise in traffic, pushing posting levels to a record high in April 2022. Coincidentally, this latter month is also the time of Slovenia's parliamentary election, one accompanied by protests and immense engagement by civil society, resulting in one of the highest voter turnouts in the history of the country.

The HBS-ijekavian collection contains $31,199,242$ tweets of which $26,975,269$ (over 86%) are original. The HBS-ekavian collection in turn contains $99,124,072$ tweets with $64,986,741$ (a bit under two thirds) original.

We observe a similar increase in tweet production during March and April 2020 as we do in the Slovenian corpus. The corpora differ from the Slovenian in that we do not observe a drop in production in subsequent months, instead we see a relatively steady growth in the amount of tweets that tapers out in 2021 and starts dropping after that. COVID-19 and measures to prevent its

spread seemed to have an effect on posting activities of HBS-ijekavian and HBS-ekavian Twitter public as well as Slovenian.

Figure 2: Number of monthly tweets in the HBS-ijekavian corpus through time.

We observe a spike in traffic during the early summer of 2020 in the HBS-ekavian corpus, which roughly aligns with parliamentary elections in Serbia. Before that, we observe a drop in production throughout the year 2019 in the HBS-ekavian corpus. Further analysis is required to fully explain it, but it is important to note that 2019 was the year of relatively large interventions into the Serbian twitter user base by Twitter itself. Consult the chart of the ratio of tweets and retweets in the following section as well as the discussion accompanying it.

Ultimately, a much deeper look into the specific months beyond the scope of this presentation of the corpora would be required to conclusively explain the spike.

3.2 How much of Twitter are echoes

By looking at the ratio between original and retweeted content, calculated as number of original tweets divided by the number of all tweets (including retweets) we can observe an interesting difference between the Slovenian, HBS-ijekavian and HBS-ekavian corpora. While the relative amount of original content is consistently dropping in the Slovenian corpus, it is actually growing

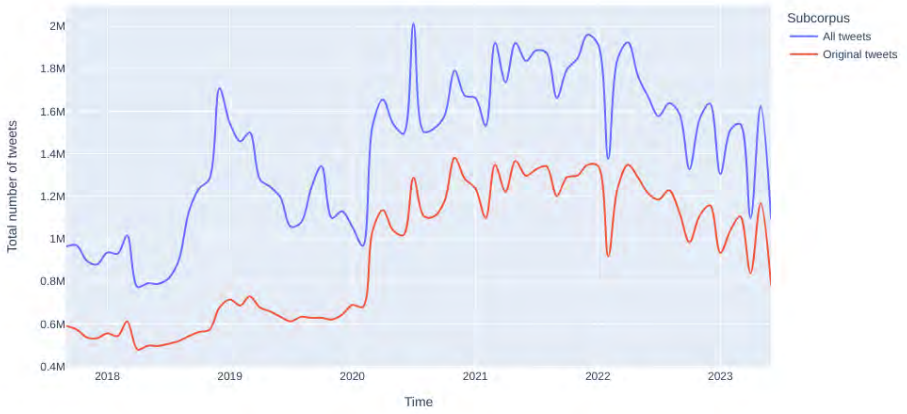

Figure 3: Number of monthly tweets in the HBS-ekavian corpus through time.

(meaning relatively fewer retweets every month) in the HBS-ijekavian and HBS-ekavian corpora. This shows a larger relative number of original content was harvested in HBS languages vs. Slovenian.

A notable drop (indicating a larger amount of retweets) occurs in the Slovenian corpus during the lockdown months of 2020 coinciding with an increase in absolute number of posts and Slovenia's 2021 parliamentary election. The ratio never recovers, instead keeps dropping with a speed similar to that before the election.

Also of note is the time between the final months of 2019 in the HBS-ekavian corpus. While the drop is relatively small, the trend of the remaining accounts continues upward. While a rigorous analysis of this dip is beyond the scope of this introductory exploration, we hypothesise it is connected with Twitter's suspension of at least 8, 558 accounts during 2019 as reported by Twitter and analysed by the Stanford Internet Observatory Cyber Policy Center (Bush, 2020).

3.3 How Twitter feels

Since all the tokens in our corpora are linguistically annotated, we can use the lemmas to draw some broad conclusions about the emotional state of Twitter. Counting each of the tokens present in each month we can then compare them

Figure 4: The ratio between tweets and original content in all three corpora.

with the LiLaH emotion lexicon (Daelemans et al., 2020) based on (Mohammad & Turney, 2010), to count the number of positively and negatively emotionally charged words. Based on those we calculate the ratio between positive and negative token counts by dividing the positive with the negative. We observe that the ratio of positive vs. negative is in favor of positively emotionally charged tokens in all three corpora. This holds true for both original tweets and those including retweets, which both show similar ratios between emotionally positive and negative tokens.

While the aforementioned LiLaH emotion lexicon only provides manual translations into Slovenian and Croatian languages, we could not find an equivalent manually translated lexicon for Serbian. The authors of the original emotion lexicon provide machine-translated lexica for other languages, among them Serbian, but using that on our HBS-ekavian corpus produced highly suspicious results. The emotional valence was effectively flatlining around 1 throughout the time we are investigating. Using the Croatian lexicon, albeit imperfect, produced much more sensible results comparatively. We are presenting the results based on the latter and strongly advise against using machine-translated lexica for similar tasks. In our particular case, the use of the machine-translated lexicon produced faulty and ultimately misleading results.

Table 2: Ratios of positive over negative token counts.

Corpus	*Original content*	*Original and retweets*
Slovenian	1.956	1.927
HBS-ijekavian	2.333	2.336
HBS-ekavian	2.404	2.406

We perform the same analysis for every month and plot it for original tweets and those with retweets included. Looking at figures 5 and 6 a notable shift towards more negative expressions can be seen in all corpora. This drop coincides with the introduction of lockdown measures from March 2020. All corpora show a negative trend in terms of tokens emotional valence signifying a slow but consistent shift in the kind of language used on Twitter.

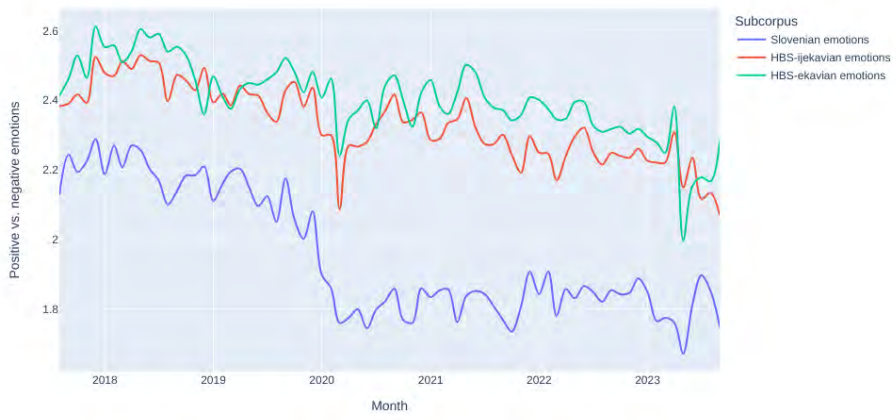

Figure 5: The ratio between positive and negative emotions in original tweets.

3.4 What Twitter talks about

There are of course multiple ways to inspect the content of tweets, but for this particular demonstration we will use the named entities recognised during our linguistic annotations. Similarly to above, we count the number of all named entities in the corpus. Listing the most common named entities in each of the person, organisation, and location categories, we can inspect the meatspace grounding of the corpora.

Figure 6: The ratio between positive and negative emotions in all tweets including retweets.

For the Slovenian corpus, the most common named people are all Slovenian politicians: Janez Janša (prime minister in during lockdown measures), Marjan Šarec(prime minister before lockdown measures), and Borut Pahor (then Slovenian president). Among organisations we find RTV Slovenija (national broadcaster), the constitutional court, and national assembly. The top three locations Novo mesto, Nova Gorica, and Murska Sobota are Slovenian cities, effectively local administrative and business hubs. We can see the corpus grounded in Slovenian national politics, media, and geography. The other two corpora do not demonstrate such national and geographic limits.

For the HBS-ijekavian corpus, we find Milorad Dodik (president of Republika Srpska, BiH), Željko Komšić (Croat member of the Presidency of BiH), and Milo Đukanović (twice president of Montenegro). Interestingly, Zoran Milanović (current president of Croatia) is the fifth most commonly named person after Novak Đoković. Looking at organisations we find the presidency of BiH, the European Union and the constitutional court. Most mentioned locations include Montenegro, BiH, and Republika Srpska. Compared to Slovenian, this corpus is much more geographically diverse, while still mostly political when it comes to recognised named entities. The results of this analysis point towards the

conclusion that a significant part of the HBS-ijekavian corpus contains tweets from Bosnia and Herzegovina.

Finally, looking at the HBS-ekavian corpus among people we find Aleksandar Vučić (president of Serbia since 2017) , Vuk Jeremić (Serbian politician), and Dragan Đilas (Serbian politician, former mayor of Belgrade). Novi Sad (Serbian city), Montenegro and Republika Srpska are the most commonly mentioned locations with Crvena Zvezda (sports club) leading among organisations, followed by the Serbian government and the party SNS (Srpska Narodna Stranka). This analaysis shows the very much expected result that in the HBS-ekavian corpus to the most part Serbian content is represented.

While smaller in the HBS-ekavian corpus, the geographical diversity of named entities even in that corpus, which would be expected to contain purely Serbian content, further supports the claim that language identities in the Balkans do not conform to national borders. Any attempt to geographically or politically locate linguistic artefacts based on the language or linguistic features is destined for failure. This should also bring into question the design, structure, and presentation of language technologies and language resources, not least the labels we use for models and their results.

3.5 What is tagged on Twitter

Another way to peek into the content of Twitter conversations is to look at the use of hashtags. Comparing most common hashtags between retweets-included and retweets-excluded collections, we can see that there is a relatively large overlap between the two perspectives. Political and news hashtags are most present, with popular culture trailing the top 20. A notable exception to the overlap is the hashtag *#Požareport* in the Slovenian corpus, which is among the top three hashtags when we include retweets, but completely absent from the originals-only subset. The hashtag is used to promote a relatively right wing news/tabloid portal previously connected with prominent Slovenian political parties. Its presence among top hashtags in the retweets-included collection only is indicative of inorganic promotion, but a deeper analysis of users and messages reproducing the tweets would be required to confirm what is usually called ”inauthentic coordinated behavior” (Cinelli et al., 2022).

Figure 7: Most popular hashtags in all Slovenian content.

Figure 8: Most popular hashtags in original Slovenian content.

Figure 9: Most popular hashtags in all HBS-ijekavian content.

Figure 10: Most popular hashtags in original HBS-ijekavian content.

Figure 11: Most popular hashtags in all HBS-ekavian content.

Figure 12: Most popular hashtags in original HBS-ekavian content.

3.6 Who talks on Twitter

The most prolific accounts in the Slovenian corpus for authors of all content (original tweets and retweets) are visualised in Figure 13. All of the top retweeters are what we can consider ”regular users”, i.e. users without an explicitly attributed organisational affiliation or publicly recognised meatspace persona. In the original content wordcloud we see the majority of users still as regular users, but only barely. 9 out of the top 20 posters of original content are either journalists or news outlets. Most prolific posters of original content are visualised as a wordcloud in Figure 14.

Figure 13: Most prolific posters of any Slovenian content.

Figure 14: Most prolific posters of original Slovenian content.

The HBS-ijekavian corpus differs from the Slovenian in that both wordclouds (posters of original content, posters of any content) contain accounts connected with news outlets. Another notable thing is that the list includes accounts which have since been deleted or suspended.

Figure 15: Most prolific posters of any HBS-ijekavian content.

Figure 16: Most prolific posters of original HBS-ijekavian content.

The HBS-ekavian corpus contains a single news outlet and one institution in the original content column. The production of HBS-ekavian tweets seems much more dominated by regular users than the other two corpora. Again, similar to the HBS-ijekavian corpus we find deleted and suspended accounts in both collections. This property of the HBS corpora potentially is consistent with Twitters reports on mass account deletions.

Figure 17: Most prolific posters of any HBS-ekavian content.

Figure 18: Most prolific posters of original HBS-ekavian content.

3.7 Who is being talked at on Twitter

Figures 19, 20, 21, 22, 23, and 24 show the most commonly mentioned accounts including and excluding retweets. We can see political accounts having the strongest presence when it comes to mentions. In this sense “private” conversations represent a vanishingly small portion of Twitter mentions. We can understand the primary use of mentions being to call out or name (and often shame) a particular politician, party, or institution. The overlap between top mentions in original and all tweets is quite large, over half of the top 20 are the same across both.

In both the HBS-ijekavian and HBS-ekavian corpora, we can see *YouTube* stand out of the political and media crowd indicating many of the captured tweets are semi-automatically generated share tweets (such as users get when using in-app share buttons). This indicates a corpus of tweets contains information about users' digital habits beyond simple online conversations.

Figure 19: Most mentioned accounts in any Slovenian content.

Figure 20: Most mentioned accounts in original Slovenian content.

Figure 21: Most mentioned accounts in any HBS-ijekavian content.

Figure 22: Most mentioned accounts in original HBS-ijekavian content.

Figure 23: Most mentioned accounts in any HBS-ekavian content.

Figure 24: Most mentioned accounts in original HBS-ekavian content.

4 CONCLUSION

In this paper we present a novel corpus of Twitter posts spanning over 6 years and 170 million tweets. We outline the data collection and processing required to obtain the corpus in its current form. We briefly discuss the split of the collection into three distinct corpora, followed by an exploratory pilot analysis of the tweets contained. We present and discuss the production of tweets through time, the relationship between the amount of original vs. retweeted content, and perform some cursory distant reading into the emotions, named entities, hashtags, authors, and mentions in the collection.

We show that the Twitter corpora created contain a wealth of information and show echoes of real life (albeit most commonly political) events. Among others, we observe the effects of COVID-19 preventative measures on the amount and quality of tweets. Furthermore, we show that while language categorisation might be useful in terms of the quality of specific linguistic annotations it would be ill-advised to use those categories to infer non-overlapping geographical or cultural categories, as shown on the HBS corpora which contain quite a lot of references to entities related to several countries, administrations, and cultural spaces.

The split of corpora presented in this paper is one of many possible, and future research would be well advised to generate subcorpora based on specific research questions it aims to address. Each of the possible splits brings with

it its own set of biases which should be evaluated at the point of use. This is especially true when researching coordinated inauthentic behaviour, signs of which are present in all three corpora, though much more noticeable in the HBS corpora (with some of the most prolific accounts from the investigated period banned today).

The differences between the three corpora presented are both qualitative and quantitative, but they serve to highlight specifics of each other. In other words, should the corpora be merged together, we would be unable to observe the spikes in tweet production during Slovenia's election. Likewise, without the split, the observed change in communication after the introduction of COVID-19 preventative measures would be visible, but the argument for its relative universality much less convincing.

The size of the collection and the results of our initial inquiry call for further research into the discourse on Twitter in the past years. While we do not have the rights to make the corpus freely available, we are retaining an archive for future text and data mining endeavours based on the exceptions provided by the European Union. We invite research teams who would like to use this dataset for their own research to contact the authors for access to the collection.

REFERENCES

Altman, M. (2018). Tufekci, z.: Twitter and tear gas: The power and fragility of networked protest. , *29*(4), 884–885. Retrieved 2024-05-31, from https://doi.org/10.1007/s11266-017-9927-0 doi: 10.1007/s11266-017-9927-0

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135–146.

Bronstein, J., Aharony, N., & Bar-Ilan, J. (2018). Politicians' use of facebook during elections: Use of emotionally-based discourse, personalization, social media engagement and vividness. doi: 10.1108/AJIM-03-2018-0067

Bush, D. (2020). "fighting like a lion for serbia": An analysis of government-linked influence operations in serbia. Retrieved 2024-05-29, from https://fsi.stanford.edu/publication/april-2020-serbia-takedown

Chan, M., & Yi, J. (2024). Social media use and political engagement in polarized times. examining the contextual roles of issue and affective polarization in developed

democracies. Routledge. Retrieved 2024-05-31, from https://www.tandfonline.com/doi/abs/10.1080/10584609.2024.2325423

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on twitter. doi: 10.1016/j.dss.2022.113819

Daelemans, W., Fišer, D., Franza, J., Kranjčić, D., Lemmens, J., Ljubešić, N., ... Popič, D. (2020). *The LiLaH emotion lexicon of croatian, dutch and slovene.* Slovenian language resource repository CLARIN.SI. http:// hdl.handle.net/11356/1318

Effing, R., van Hillegersberg, J., & Huibers, T. (2011). Social media and political participation: Are facebook, twitter and YouTube democratizing our political systems? In E. Tambouris, A. Macintosh, & H. de Bruijn (Eds.), *Electronic participation* (pp. 25–35). Springer. doi: 10.1007/978-3-642-23333-3_3

Flew, T., & Iosifidis, P. (2020). Populism, globalisation and social media. , *82*(1), 7–25. SAGE Publications Ltd. Retrieved 2024-05-31, from https://doi.org/10.1177/1748048519880721 doi: 10.1177/1748048519880721

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. , *8*, (pp.67698–67717). IEEE Access Retrieved 2024-05-31, from https://ieeexplore.ieee.org/document/9047963 doi: 10.1109/ACCESS.2020.2983656

Labrèche, G. (2023, March). *Cyrtranslit.* Zenodo. (A Python package for bi-directional transliteration of Cyrillic script to Latin script and vice versa. Supports transliteration for Bulgarian, Montenegrin, Macedonian, Mongolian, Russian, Serbian, Tajik, and Ukrainian.) doi: 10.5281/ zenodo.7734906

Ljubešić, N., Fišer, D., & Erjavec, T. (2014). TweetCaT: a tool for building Twitter corpora of smaller languages. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 2279–2283). Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf

Ljubešić, N., & Kranjčić, D. (2014). Discriminating between very similar languages among twitter users. In *Proceedings of the ninth language technologies conference* (pp. 90–94).

Ljubešić, N., Petrović, M. M., & Samardžić, T. (2018). Borders and boundaries in bosnian, croatian, montenegrin and serbian: Twitter data to the rescue. *Journal of Linguistic Geography*, *6*(2), 100–124.

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and genera-*

tion of emotion in text (pp. 26–34). Los Angeles, CA: Association for Computational Linguistics. https://aclanthology.org/W10-0204

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages.

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using twitter data: a comparative application of lexicon- and machine-learning-based approach. , *13*(1), 31. Retrieved 2024-05-31, from https://doi.org/10.1007/s13278-023-01030-x doi: 10.1007/s13278-023-01030-x

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data. *7*(1), 13006. Retrieved 2024-05-31, from https://www.nature.com/articles/s41598-017-12961-9 doi: 10.1038/s41598-017-12961-9

Rupnik, P., Kuzman, T., & Ljubešić, N. (2023). BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, & M. Zampieri (Eds.), *Tenth workshop on nlp for similar languages, varieties and dialects (vardial 2023)* (pp. 113–120). Dubrovnik, Croatia: Association for Computational Linguistics. https://aclanthology.org/2023.vardial-1.11 doi: 10.18653/v1/2023.vardial-1.11

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The next step for linguistic processing of south slavic languages.

Wang, L., Niu, J., & Yu, S. (2020). SentiDiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis. In: IEEE Transactions on knowledge and data engenireenig, *32*(10), 2026–2039. Retrieved 2024-05-31, from https://ieeexplore.ieee.org/document/8700266 doi: 10.1109/TKDE.2019.2913641

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In P. Clough et al. (Eds.), *Advances in information retrieval* (pp. 338–349). Springer. doi: 10.1007/978-3-642-20161-5_34

TWITTER PRED X: PODOBE Z BALKANA

Predstavljamo nov korpus preko 170 milijonov tvitov v Slovenskem in jezikih HBS, zbranih med letoma 2017 in 2023. Po opisu procesa zbiranja in (pred)procesiranja podatkov predstavimo odločitev o razdelitvi korpusa v tri podkorpuse, vključno z ovirami na poti jezikovnega označevanja zbirke iz tako narodnostno in jezikovno mešanega področja kot je Balkan. Nadaljujemo s pilotno analizo kvalitete, količine in razlik med zbirkami. Predstavimo indice o vplivu dogodkov v resničnem življenju na uporabniško generirane vsebine na spletu. Po predstavitvi nekaterih najbolj vidnih vsebinskih lastnosti zbirke zaključimo z vabilom k dodatnemu raziskovanju tega obsežnega in raznolikega korpusa sodobnega spletnega govora.

Keywords: Twitter korpus, slovenščina, makro jezik HBS, oddaljeno branje.