EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+

Matej Martinc
Jožef Stefan Institute
Jamova 39, Ljubljana
matej.martinc@ijs.si

Nina Perger
Faculty of Social Sciences
Kardeljeva ploščad 5, Ljubljana
nina.perger@fdv.uni-lj.si

Andraž Pelicon
Jožef Stefan Institute
Jamova 39, Ljubljana
andraz.pelicon@@ijs.si

Matej Ulčar
Faculty of Computer Science
Večna pot 113, Ljubljana
matej.ulcar@fri.uni-lj.si

Andreja Vezovnik
Faculty of Social Sciences
Kardeljeva ploščad 5, Ljubljana
andreja.vezovnik@fdv.uni-lj.si

Senja Pollak
Jožef Stefan Institute
Jamova 39, Ljubljana
senja.pollak@ijs.si

Abstract

We conduct automatic sentiment and viewpoint analysis of the newly created Slovenian news corpus containing articles related to the topic of LGBTIQ+ by employing the state-of-the-art news sentiment classifier and a system for semantic change detection. The focus is on the differences in reporting between quality news media with long tradition and news media with financial and political connections to SDS, a Slovene right-wing political party. The results suggest that political affiliation of the media can affect the sentiment distribution of articles and the framing of specific LGBTIQ+ specific topics, such as same-sex marriage.

1 Introduction

Quantitative content analysis of news related to LGBTIQ+ in general, and specifically, to marriage equality debates show that distinctions can be drawn between those media articles that express positive, neutral or negative stance towards same-sex marriage. Those media articles that express positive stance are grounded in human rights/civil equality discourses and access to benefits (Zheng and Chan, 2020; Colistra and Johnson, 2019; Paterson and Coffey-Glover, 2018), and frame marriage equality as an inevitable path towards equality, as a civil right issue that would reduce existing prejudices and discrimination, and protect threatened LGBTIQ+ minority (Zheng and Chan, 2020).

For media articles that express negative stance towards marriage equality, distinctive discursive elements are present, such as "equal, but separate" (marriage equality should be implemented, but differentiating labels should be kept in the name of protecting the institute of marriage) (Kania, 2020; Zheng and Chan, 2020; Paterson and Coffey-Glover, 2018), and reference procreation/welfare of children (Kania, 2020; Zheng and Chan, 2020), public objection (Kania, 2020) and church – state opposition (Paterson and Coffey-Glover, 2018).

The related work also shows that the differences between "liberal" and "conservative" arguments are not emphasised, mostly because both sides refer to each other's arguments, if only to negate them; yet, political orientation can be identified through the tone of the article (Zheng and Chan, 2020).

When it comes to methods employed for automatic analysis of the LGBTIQ+ topic, most recent approaches rely on embeddings. Hamilton et al. (2016) employed embeddings to research how words (among them also word *gay*) change meaning through time. They built static embedding models for each time slice of the corpus and then make these representations comparable by employing *vector space alignment* by optimising a geometric transformation. This research was recently expanded by (Shi and Lei, 2020), who employed embeddings to explore semantic shifts of six descriptive LGBTIQ+ words from the 1860s to the 2000s: *homosexual, lesbian, gay, bisexual, transgender*, and *queer*.

There are also several general news analysis techniques that can be employed for the task at hand. Azarbonyad et al. (2017) developed a system for semantic shift detection for viewpoint analysis of political and media discourse. A recent study by Spinde et al. (2021) tried to identify biased terms in news articles by comparing news media outlet specific word embeddings. On the other hand, Pelicon et al. (2020) developed a system for analysing the sentiment of news media articles.

While the above described analyses in a large majority of cases covered news in English speaking countries, in this research, we expand the quantitative analysis to Slovenian news, in order to determine whether attitudes towards LGBTIQ+ differs in different cultural environments. We created a corpus of LGBTIQ+ related news and conducted an

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 121–126
April 19, 2021

automatic analysis of its content covering several aspects:

- Sentiment of news reporting, where we focused on the differences in reporting between well established media with long tradition of news reporting and more recently established media characterised by their financial and political connections to the Slovene conservative political party SDS.
- Usage of words, where we tried to identify the words that are used differently in different news sources and would indicate the difference in the prevailing discourse on the topic of LGBTIQ+ in the specific liberal and conservative media.

The research was performed in the scope of the EMBEDDIA Hackashop (Hackaton track) at EACL 2021 and employs several of the proposed resources and tools (Pollak et al., 2021).

2 Methodology

For **sentiment analysis** we used a multilingual news sentiment analysis tool. The tool was trained using a two-step approach, described in Pelicon et al. (2020). For training, a corpus of sentiment-labeled news articles in Slovenian was used (Bucar et al., 2018) with news covering predominantly the financial and political domains. This model was subsequently applied to the LGBTIQ+ corpus where each news article was labeled with one of the sentiment labels, namely negative, neutral or positive. This allowed us to generate a sentiment distribution of articles for each media source in the corpus.

For **word usage viewpoints analysis**, we applied a system originally employed for diachronic shift detection (Martinc et al., 2020b). Our word usage detection pipeline follows the procedure proposed in the previous work (Martinc et al., 2020a,b; Giulianelli et al., 2020): the created LGBTIQ+ corpus is split into two slices containing news from different news source according to procedure described in Section 3. Next, the corpus is lemmatized, using the Stanza library (Qi et al., 2020), and lowercased. For each lemma that appears more than 100 times in each slice and is not considered a stopword, we generate a slice specific set of contextual embeddings using BERT (Devlin et al., 2019) pretrained on the Slovenian, Croatian and English texts (Ulčar and Robnik-Šikonja, 2020). These representations are clustered using k-means and the derived cluster distributions are compared across slices by employing Wasserstein distance (Solomon, 2018). It is assumed that the ranking resembles a relative degree of usage change, therefore words are ranked according to the distance.

Once the most changed words are identified, the next step is to understand how their usage differs in the distinct corpus slices. The hypothesis is that specific clusters of BERT embeddings resemble specific word usages of a specific word. The problem is that these clusters may consist of several hundreds or even thousands of word usages, i.e. sentences, therefore manual inspection of these usages would be time-consuming. For this reason, we extract the most discriminating unigrams, bigrams, trigrams and fourgrams for each cluster using the following procedure: we compute the term frequency - inverse document frequency (tf-idf) score of each n-gram and the n-grams appearing in more than 80% of the clusters are excluded to ensure that the selected keywords are the most discriminant. This gives us a ranked list of keywords for each cluster and the top-ranked keywords (according to tf-idf) are used for the interpretation of the cluster.

3 Experiments

3.1 Dataset

The corpus was collected from the Event registry (Leban et al., 2014) dataset by searching for Slovenian articles from 2014 to (including) 2020, containing any of the manually defined 125 keywords (83 unigrams and 42 bigrams) and their inflected forms connected to the subject of LGBTIQ+. The resulting corpus contains news articles on the LGBTIQ+ topic from 23 media sources. The corpus statistics are described in Table 1. Out of this corpus, we extracted a subcorpus appropriate for the viewpoint analysis. The subcorpus we used included the following online news media: Delo, Večer, Dnevnik, Nova24TV, Tednik Demokracija and PortalPolitikis. The sources were divided into two groups. The first group, namely Delo, Večer and Dnevnik represent the category of daily quality news media that are published online and in print with a long tradition in the Slovene media landscape. These three media are relatively highly trusted by readers and have the highest readership amongst Slovene dailies. The second group of news media - namely, Nova24TV, Ted-

Source	Num. articles	Num. words
MMC RTV Slovenija	1790	1,555,977
Delo	1194	1,064,615
Nova24TV	844	683,336
Večer	667	552,195
24ur.com	661	313,794
Dnevnik	592	262,482
Siol.net Novice	549	460,561
Slovenske novice	501	236,516
Svet24	430	286,429
Mladina	394	275,506
Tednik Demokracija	361	350,742
Domovina	327	283,478
Primorske novice	255	183,624
Druzina.si	253	149,761
Vestnik	242	263,737
Časnik.si - Spletni magazin z mero	239	280,339
Žurnal24	172	79,953
PortalPolitikis	157	111,683
Revija Reporter	102	62,429
Gorenjski Glas	97	92,751
Onaplus	79	104,343
Športni Dnevnik Ekipa	67	33,936
Cosmopolitan Slovenija	57	71,538

Table 1: LGBTIQ+ corpus statistics.

nik Demokracija and PortalPolitikis have been established more recently and are characterised by their financial and political connections to the Slovene right-wing/conservative political party SDS (Slovenska demokratska stranka) and the Roman Catholic Church.

3.2 Sentiment Analysis

Figure 1 presents sentiment distribution across articles for each specific news media, arranged from left to right according to the share of articles with negative sentiment. Note that all three media houses selected for the viewpoint analysis (Nova24TV, Tednik, Demokracija and PortalPolitikis) because of their financial and political connections to the Slovene right-wing/conservative political party SDS produce more news articles with negative sentiment on the topic of LGBTIQ+ than the mainstream media with the long tradition (Delo, Dnevnik, Večer). The source with the most negative content about LGBTIQ+ is Revija Reporter, which is in most media analyses positioned in the right-wing ideological spectrum[1] (Milosavljević, 2016; Milosavljević and Biljak Gerjevič, 2020). On the other side the source with the smallest share of negative news is Primorske novice, a politically independent daily regional quality news media published online and in print with a long tradition in the regional media landscape. Nevertheless, not all conservative media are characterized by a more negative reporting about the LGBTIQ+ topic. For example, the source with the second lowest share of negative news is Druzina.si, which is strongly connected to Roman Catholic Church.

3.3 Viewpoint Analysis

The viewpoint analysis was conducted by finding words, whose usage varies the most in the two groups of media sources selected for the analysis (i.e. Delo, Dnevnik, Večer vs. Nova24TV, Tednik Demokracija and PortalPolitiks). The 10 most changed words are presented in Table 2. The word that changed the most was globok (deep), for which our system for interpretation of the change revealed that it was selected due to frequent mentions of *deep state* in the media with connections to political right. The context of *deep state* is interesting, since it is a very frequently used interpretative frame by this group of media sources, regardless of the specific topic. Here it indicates the framing of the LGBTIQ+ questions as part of a political agenda driven by the left-wing politics. The second word roman (novel) was selected because it appears in two contexts: as a novel and also as a constituent word in a name of the Slovenian LGBTIQ+ activist, Roman Kuhar. While the third word, *video*, is a corpus artefact that offers little insight into the attitude towards LGBTIQ+, the fourth word, *razmerje* (relationship), has a direct connection to some of the most dividing LGBTIQ+ topics, such as gay marriage, therefore for this word we provide a more detailed analysis. Figure 2 presents cluster distributions per two media groups and top 5 (translated) keywords for each cluster for word *razmerje(relationship)*. The main difference between the two distributions can be observed when it comes to mention of relationship in the context of family and marriage (see the red cluster), which present a large cluster of usages in the mainstream media but a rather small cluster in the right-wing

1	globok(deep)	6	napaka(mistake)
2	roman(novel)	7	nadaljevanje(continuation)
3	video	8	lanski(last year)
4	razmerje(relationship)	9	kriza(crisis)
5	teorija(theory)	10	pogledat(look)

Table 2: Top 10 most changed words (and their English translations) in the corpus according to Wasserstein distance between k-means ($k = 5$) cluster distributions in distinct chunks of the corpus.

[1] https://podcrto.si/mediji-martina-odlazka-1-del-nepregledna-mreza-radiev-tiskovin-televizije/

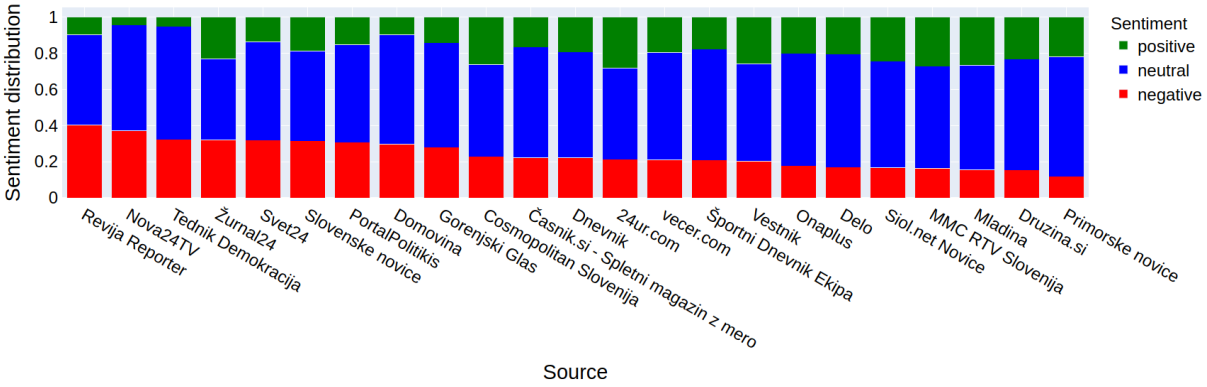
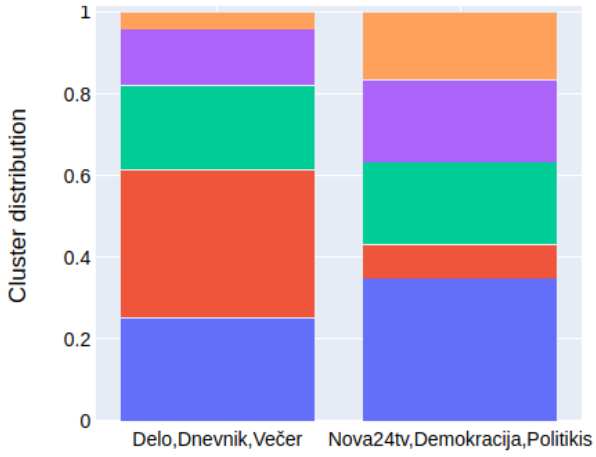


Figure 1: Sentiment distribution for each source in the LGBTIQ+ corpus.



#	Keywords
	partnership, homosexual relationship, polygamous relationship, intimate relationship, problem
	family relationship, end work relationship, woman, field, less
	power in the relationship, relationship between people, percentage, man and woman, story
	family relationship, marriage, children, Ljubljana, coalition for children
	woman, love relationship, work relationship, film, meet

Figure 2: Cluster distributions per two media groups and top 5 translated keywords for each cluster for word *razmerje(relationship)*.

media. On the other hand, relationship is in these media mentioned a lot more in the context of partnership, homosexuality and polygamy (see the orange cluster). The other three clusters (i.e., usages) have a rather strong presence in both media groups.

4 Conclusions

We conducted a content analysis of the Slovenian news corpus containing articles related to the topic of LGBTIQ+. The sentiment analysis study shows that there are some differences in the sentiment of reporting about LGBTIQ+ between two distinct groups of media and that the three media houses connected to political right tend to cover the subject in a more negative manner. This supports the thesis by Zheng and Chan (2020), who suggested that political orientation can be identified through the tone of the article. Nevertheless, the obtained results should be interpreted with the grain of caution, since the sentiment classifier we employed cannot distinguish whether it is the stance expressed towards the LGBTIQ+ community, or is it rather the event on which the article is reporting, that is positive or negative (e.g., an attack on the LGBTIQ+ activist). The distinction between these two "types" of sentiment will be analysed in the future work.

The viewpoint analysis suggests that the usage of some specific words has been adapted in order to express specific ideological point of view of the media. For example, the analysis of the word *relationship* suggests that the more conservative media more likely frame LGBTIQ+ relationships as a *partnership* of two homosexual (or even polygamous) partners. On the other hand, they rarely consider LGBTIQ+ relationships as family or talk about marriage.

In the future we plan to conduct topic analysis of the corpus in order to identify the most common LGBTIQ+ related topics covered by the news media. We will also employ embeddings to research relations between LGBTIQ+ specific words.

Acknowledgments

This work was supported by the Slovenian Research Agency (ARRS) grants for the core programme Knowledge technologies (P2-0103), the project Computer-assisted multilingual news discourse analysis with contextual embeddings (CANDAS, J6-2581), as well as the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518.

Joze Bucar, M. Znidarsic, and J. Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52:895–919.

Rita Colistra and Chelsea Betts Johnson. 2019. Framing the legalization of marriage for same-sex couples: An examination of news coverage surrounding the us supreme court's landmark decision. *Journal of homosexuality*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501.

Ursula Kania. 2020. Marriage for all ('ehe fuer alle')?! a corpus-assisted discourse analysis of the marriage equality debate in germany. *Critical Discourse Studies*, 17(2):138–155.

Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 107–110.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020a. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA. Association for Computing Machinery.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Discovery team at semeval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73.

Marko Milosavljević. 2016. Media pluralism monitor 2016 monitoring risks for media pluralism in the EU and beyond - country report: Slovenia.

Marko Milosavljević and Romana Biljak Gerjevič. 2020. Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the European Union, Albania and Turkey in the years 2018-2019 - country report: Slovenia.

Laura L Paterson and Laura Coffey-Glover. 2018. Discourses of marriage in same-sex marriage debates in the uk press 2011–2014. *Journal of Language and Sexuality*, 7(2):175–204.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.

Senja Pollak, Marko Robnik Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Yaqian Shi and Lei Lei. 2020. The evolution of lgbt labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes. *English Today*, 36(4):33–39.

Justin Solomon. 2018. Optimal transport on discrete domains.

Timo Spinde, Lada Rudnitckaia, and Felix Hamborg. 2021. Identification of biased terms in news articles

by comparison of outlet-specific word embeddings. In *Proceedings of the 16th International Conference (iConference 2021). Springer Nature, Virtual Event China*.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert: less is more in multilingual models. *arXiv preprint arXiv:2006.07890*.

Yue Zheng and Lik Sam Chan. 2020. Framing same-sex marriage in us liberal and conservative newspapers from 2004 to 2016: Changes in issue attributes, organizing themes, and story tones. *The Social Science Journal*, pages 1–13.