Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+

Matej Martinc[1] Nina Perger[2]
Andraž Pelicon[1] Matej Ulčar[3]
Andreja Vezovnik[2] Senja Pollak[1]

(1) Jozef Stefan Institute
(2) University of Ljubljana - Faculty of social sciences
(3) University of Ljubljana - Faculty of computer and information science

EMBEDDIA presentation Paris - November 8^{th} 2021

What was the main purpose of the research?

Media representation of specific issues is important.

Some issues related to LGBTIQ+ are dividing the public opinion.

Content analysis of news in English speaking countries showed that **distinctions can be drawn between articles that express different stances towards same-sex marriage**:

- Media articles that express **positive** stance mostly refer to **civil equality/human rights** [Zheng and Chan, 2020].
- In media articles that express **negative stance**, **"equal, but separate"** discourse is present [Zheng and Chan, 2020].

No such quantitative analysis has yet been conducted on Slovenian news

What are the main differences in reporting about LGBTIQ+ in different Slovenian media?

Approach

- **Build a corpus** of LGBTIQ+ news from many media sources
- **Conduct sentiment and viewpoint analysis** on the corpus

Which tools did we use and how?

Sentiment analysis

We used a multilingual news sentiment analysis tool described in [Pelicon et al., 2020].

- **For training, a corpus of sentiment-labeled news articles** in Slovenian was used [Bucar et al., 2018].
- This model was applied to the LGBTIQ+ corpus where each news article was **labeled with negative, neutral or positive label**.
- This allowed us to generate a **sentiment distribution of articles** for each media source in the corpus.

Viewpoint analysis

We conducted a word usage viewpoint analysis, employing a system originally employed for diachronic shift detection [Martinc et al., 2020].

The procedure was the following:

1. The LGBTIQ+ corpus is **split into two slices**.
2. The corpus is **lemmatized** and **lowercased**.
3. For each lemma we generate a **slice specific set of contextual embeddings using BERT** pretrained on the Slovenian, Croatian and English texts [Ulčar and Robnik-Šikonja, 2020].
4. **Representations are clustered using k-means** and the derived cluster distributions are **compared across slices by employing Wasserstein distance**.
5. Words are ranked according to the distance ⇒ distance indicates usage change!

Interpretation

- The hypothesis is that specific clusters of BERT embeddings resemble specific usages of a word.
- We treat this clusters as documents and weight unigrams, bigrams, trigrams and fourgrams in the corpus with tf-idf.
- This gives us a **ranked list of keywords for each cluster** and the top-ranked keywords are used for the interpretation of the cluster.

Experiments

Corpus creation

lgbt	lgbtq	lgbtiq	lgbtiq+
lgbt ideologija	lgbt lobi	lgbt agenda	homoseksualnost
homoseksualen	homoseksualna	homoseksualno	seksualna identiteta
spolna usmerjenost	spolno usmerjen	spolno usmerjena	spolno usmerjeno
seksualna usmerjenost	seksualno usmerjen	seksualno usmerjena	seksualno usmerjeno
istospolna privlačnost	spolna perverzija	seksualna perverzija	lezbičnost
lezbištvo	lezbijka	lezba	lezbača
homoseksualka	lezbičen	lezbična	lezbično
gejevstvo	istospolen	istospolna	istospolno
gej	peder	sodomit	toplovodar
pederast	buzerant	buzerantski	buzerantska
buzerantsko	homoseksualec	gejevski	gejevska
gejevsko	pederski	pederska	pedersko
biseksualnost	biseksualka	biseksualec	biseksualen
biseksualna	biseksualno	panseksualnost	panseksualka
panseksualec	panseksualen	panseksualna	panseksualno
aseksualnost	aseksualka	aseksualec	aseksualen
aseksualna	aseksualno	kvir	queer
kvirovski	kvirovska	kvirovsko	queerovski
queerovska	queerovsko	transspolnost	spolna tranzicija
sprememba spola	tranzicija spola	potrditev spola	priznanje spola
biološki spol	spolna disforija	spolno disforičen	spolno disforična
spolno disforično	tretji spol	teorija spola	ideologija spola
transeksualnost	transseksualka	transseksualec	transvestit
transvestitka	transspolno	transspolna	transspolen
transspolnik	transspolnica	spolna identiteta	trans ženska
trans moški	transspolna ženska	transspolni moški	spolna nebinarnost
spolno nebinaren	spolno nebinarna	spolno nebinarno	spolna fluidnost
spolno fluiden	spolno fluidna	spolno fluidno	izbira spola
interseksualnost	interspolnost	interspolno	interspolna
interspolen	hermafrodit	hermafroditka	obojespolnik
obojespolnica			

Table: LGBTQ+ keywords used to extract the corpus from the Event Registry dataset.

Corpus structure

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

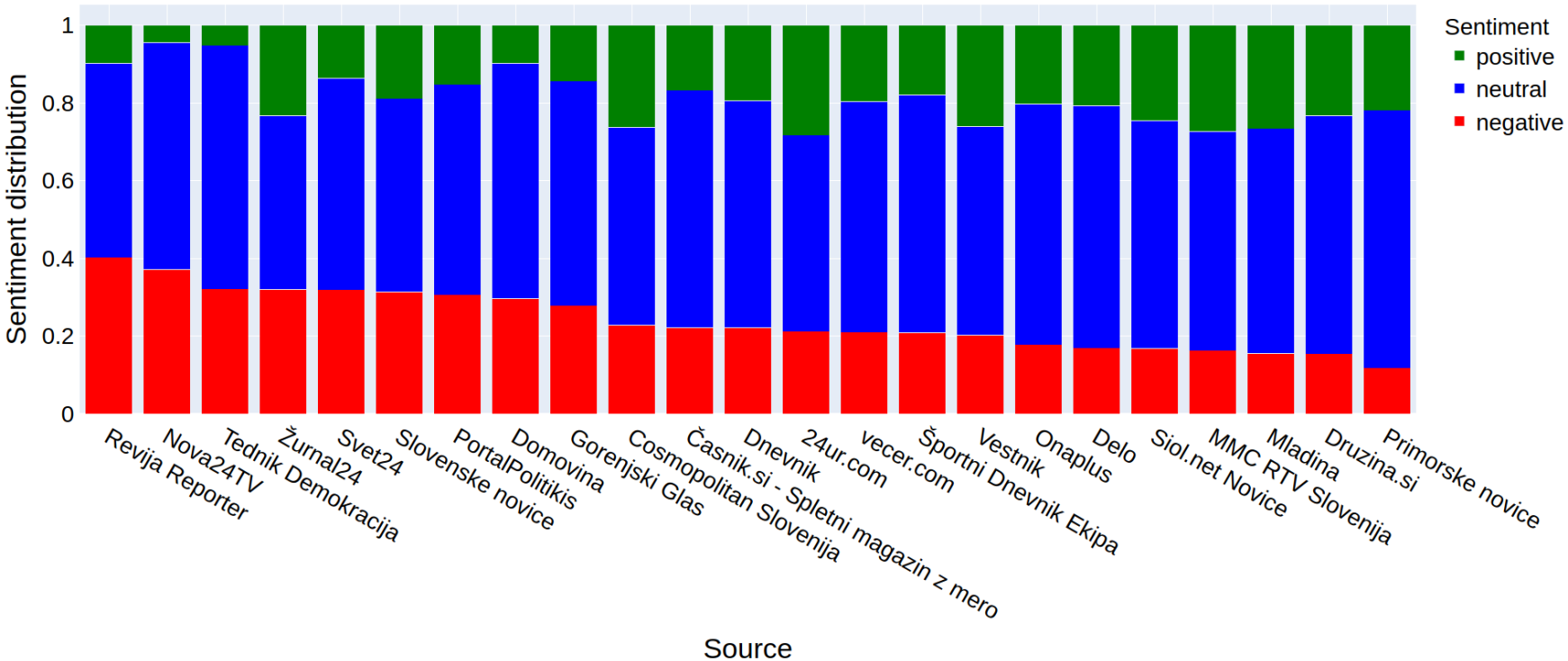
Table: LGBTQ+ corpus statistics.

Viewpoint division

We only used a subcorpus for viewpoint detection, which included the following media:

- Delo, Večer and Dnevnik represent the category of daily quality news media with a long tradition in the Slovene media landscape. These three media **have the highest readership amongst Slovene daily newspapers**.
- Nova24TV, Tednik 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 and the Roman Catholic Church**.

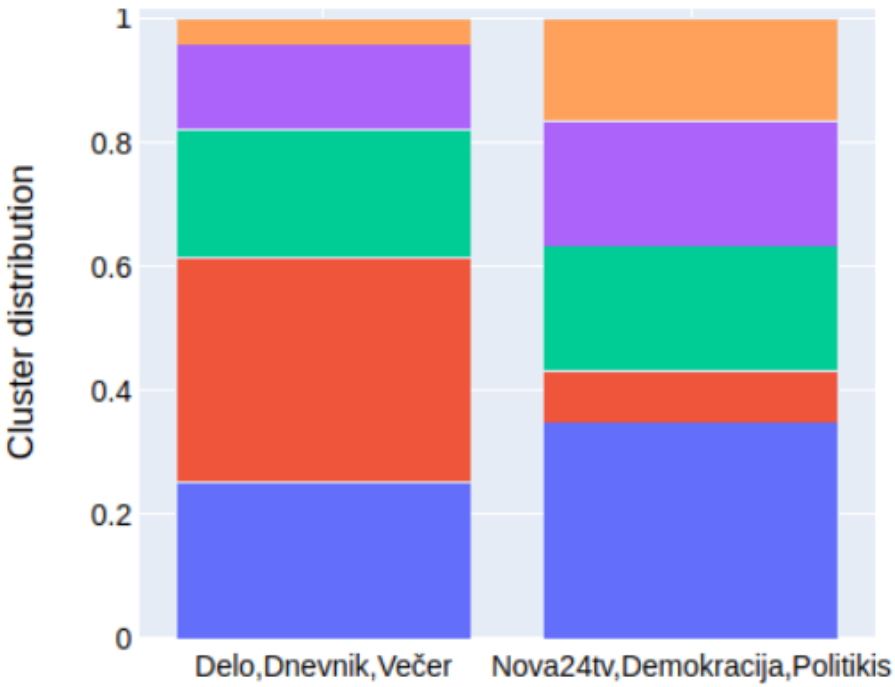
Sentiment distribution across news media sources



Viewpoint analysis

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: Top 10 most changed words (and their English translations) in the 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: Cluster distributions per two media groups and top 5 translated keywords for each cluster for word *razmerje(relationship)*.

Conclusion

Conclusion

- **The three media houses connected to political right tend to cover the LGBTIQ+ 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.
- 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.**
- 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**.

Thank you for your attention!

Question?

References

References I

Bucar, J., Znidarsic, M., and Povh, J. (2018).

Annotated news corpora and a lexicon for sentiment analysis in slovene.

Language Resources and Evaluation, 52:895–919.

Martinc, M., Montariol, S., Zosa, E., and Pivovarova, L. (2020).

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.

Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., and Pollak, S. (2020).

Zero-shot learning for cross-lingual news sentiment classification.

Applied Sciences, 10(17):5993.

Ulčar, M. and Robnik-Šikonja, M. (2020).

Finest bert and crosloengual bert: less is more in multilingual models.

arXiv preprint arXiv:2006.07890.

Zheng, Y. and Chan, L. S. (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.