University of Bern, Fall Semester 2023

Department of social sciences

Computational Social Science - Accessing and using digital data and technologies

KSL-No.: 456027

Seminar Paper

Swiss media coverage of the refugee issue: a sentiment analysis

Dr. Rudolf Farys and PD Dr. Achim Edelmann

Bern, 1/9/2024

Simon Bernhard

Undergraduate student

Social Sciences, sem. 5

simon.bernhard@students.unibe.ch

21-126-461

Kevin Jan Schläpfer

Undergraduate student

Social Sciences, sem. 5

kevin.schlaepfer@students.unibe.ch

21-125-034

Table of contents

Table of Figures

List of Tables

Introduction

One of the main political issues in Europe over the last decade has been how to deal with refugees. This debate has intensified with the refugee crisis in 2015 and with the outbreak of war in the Ukraine in February of 2022. In Switzerland namely the Swiss People's Party (SVP) has put the issue of refugees and migrants on the political agenda (e.g. SRF 2015). Accordingly to the political agenda, this is an important topic on the media agenda as well.

This paper examines changes in the media coverage of refugees in Switzerland has changed over the last decade as this is an important influence on public opinion regarding this issue. This article tries to answer the following research question: *How does the tonality of the media regarding asylum migration in Switzerland changes over time and in comparison to absolute immigration figures?*

By collecting a large dataset of articles on the topic of refugees and asylum migration from the largest German-speaking Swiss media and applying a sentiment analysis, we extracted the media's tonality when reporting about asylum migration.

This paper is structured as follows: First we describe our data source and the dataset used. Then we document the pre-processing, cleaning and analysis steps applied. In the *results* chapter we present our outcomes and interpret them. Lastly, we conclude this paper with a discussion of our results.

Data and methods

The data source for this study was the tool Swissdox@LirI. This tool has over 29 million articles from different Swiss media and gets updated daily (Swissdox AG n.d.). By using the API interface offered by the Swissdox@LirI tool we retrieved a dataset with 188'196 articles. This dataset contains articles in the period from the year 2010 to the year 2022 which had certain key word stems concerning asylum migration in the text. The root words were the German words for refugees, asylum, immigration, to flee and to immigrate ("Flücht*", "Asyl*", "Immig*", "flücht*", "asyl*", "immig*" [1]). The dataset is limited to German-speaking media and contains all relevant media outlets with a big circulation (according to BfS 2023). Included are articles from the online and print versions of 20 Minuten, Berner Zeitung, Blick am Abend, der Bund, Neue Zürcher Zeitung, Tagesanzeiger, Aargauer Zeitung, Luzerner Zeitung and Weltwoche. Also included are the print media Basler Zeitung, St. Galler Tagblatt, Schweizer

[1] The asteriks are used as placeholder to match all different expressions with this word stem.

Woche, Sonntagsblick and Die Wochenzeitung and the online media swissinfo.ch, srf.ch and weltwoche-daily.ch.

Upon inspecting the raw dataset, it became clear, that the search scope was too broad as articles concerning different issues than asylum migration were included. Therefore, we decided to utilize the unsupervised machine learning method Latent Dirichlet Allocation (LDA) to assign each article to a topic as this was the only feasible method to filter the articles. For this purpose, we cleaned the text fields from JavaScript remnants and constructed a corpus with the important variables. Then the corpus was transformed to tokens and underwent some data tidying steps, including the removal of stop words and stemming of the text. The German stop words dictionary within the quanteda package was used to remove stop words. As the goal was to get clear and distinguishable topics, we added some custom stop words which we retrieved from test running the LDA model. As the test running of the LDA model showed various single character tokens, we removed them using a regular expression. Finally, empty tokens were removed and all tokens were converted to lowercase.

The cleaned tokens were transformed into a document-feature matrix (dfm). We fitted a LDA model with 10 topics to the dfm and found that 10 was the optimal number of topics as it returned the most distinct topics. Figure 1 displays 20 terms for each topic of the LDA model. Each article was assigned one of the ten topics, and we only kept the articles with the topics 3, 7 and 10, as these are the topics related to asylum migration. Because the LDA cannot be reproduced, as it calculates different outcomes every time it is run, we included our used LDA model in the appendix.

	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10
[1,]	"bild"	"schweiz"	"ukrain"	"st.gall"	"schweiz"	"jahr"	"kanton"	"prozent"	"mann"	"fluchtling"
[2,]	"jahr"	"polit"	"russisch"	"bild"	"jahr"	"imm"	"jahr"	"jahr"	"polizei"	"land"
[3,]	"person"	"jahr"	"russland"	"septemb"	"erst"	"leb"	"fluchtling"	"schweiz"	"tat"	"mensch"
[4,]	"neu"	"partei"	"sagt"	"jahr"	"spiel"	"mensch"	"schweiz"	"frank"	"kantonspolizei"	"jahr"
[5,]	"novemb"	"bundesrat"	"kiew"	"neu"	"de"	"frau"	"gemeind"	"million"	"zurich"	"europa"
[6,]	"appenzell"	"svp"	"krieg"	"stadt"	"letzt"	"gut"	"bern"	"unternehm"	"frau"	"regier"
[7,]	"thurgau"	"muss"	"putin"	"novemb"	"zweit"	"viel"	"asylsuch"	"milliard"	"verletzt"	"migrant"
[8,]	"gemeind"	"neu"	"angab"	"samstag"	"saison"	"heut"	"kind"	"neu"	"konnt"	"viel"
[9,]	"septemb"	"wahl"	"land"	"oktob"	"minut"	"gross"	"person"	"hoh"	"person"	"israel"
[10,]	"oktob"	"gross"	"prasident"	"kanton"	"sieg"	"welt"	"stadt"	"erst"	"weg"	"grenz"
[11,]	"kanton"	"land"	"gebiet"	"mitteil"	"platz"	"bild"	"viel"	"zahl"	"mehr"	"sagt"
[12,]	"frank"	"frag"	"stadt"	"schweiz"	"team"	"macht"	"zurich"	"vergang"	"jahr"	"turkei"
[13,]	"ausserrhod"	"war"	"selenski"	"erst"	"neu"	"jung"	"frank"	"mitteil"	"polizist"	"polit"
[14,]	"prozent"	"gut"	"angriff"	"freitag"	"het"	"mann"	"mensch"	"bereit"	"staatsanwaltschaft"	"staat"
[15,]	"kantonspolizei"	"europa"	"moskau"	"ausstell"	"leagu"	"neu"	"muss"	"heisst"	"beid"	"deutschland"
[16,]	"konnt"	"grun"	"mensch"	"findet"	"wm"	"lang"	"gut"	"umsatz"	"fluchtet"	"usa"
[17,]	"erst"	"recht"	"trupp"	"donnerstag"	"train"	"geschicht"	"neu"	"wenig"	"opfer"	"prasident"
[18,]	"weit"	"nationalrat"	"weit"	"august"	"isch"	"kind"	"bund"	"deutlich"	"unbekannt"	"syri"
[19,]	"schweiz"	"prozent"	"soldat"	"person"	"sekund"	"weiss"	"famili"	"gegenub"	"kam"	"lag"
[20,]	"mitteil"	"fall"	"region"	"sonntag"	"final"	"geht"	"arbeit"	"gut"	"weit"	"eu"

Figure 1 20 terms of the LDA topics (own illustration)

Because the dataset was changed (notably by removing the custom stop words) it wasn't possible to use this dataset for a sentiment analysis. Therefore, we performed a filtering join of the original dataset and the filtered one to obtain a dataset containing the original versions of the articles related to the topic of asylum migration. During this process some articles with the

same ID's were removed to avoid interference with the joining process. Upon further investigation it was discovered that the majority of the articles with identical ID's were live ticker articles or articles that were frequently updated such as summary articles to the current state of the Ukraine war. After the whole process, the dataset was reduced to 75'169 articles.

The tonality of the media articles was analyzed using the method of sentiment analysis. The filtered dataset was prepared for sentiment analysis using the same pre-processing steps as for the LDA. To perform the sentiment analysis, the SentiMerge dictionary, which combines four different German sentiment dictionaries, was used (Emerson and Declerck 2014). The dictionary underwent some tidying steps and was then applied to the dfm of the filtered articles. The calculated sentiment scores for each article were then z-standardized.

To determine whether the tonality of the media articles correlates with the actual immigration data in Switzerland, we calculated two linear regression models. For this purpose we used monthly asylum statistics collected in Switzerland (see SEM 2023) as the independent variable and sentiment scores of the articles as the dependent variable in both models. The first model utilized the monthly total of immigration numbers as the independent variable, while the second model used the origin continents of the immigrants as the independent variable. Because the asylum statistics were only available from 2013 to 2022, both regression models and other visualizations including the asylum statistics range only from 2013 to 2022.

Results

Descriptive statistics and visualizations

The articles' sentiment scores ranged from -6.6 to 5.5 with a median of 0.028 and a mean of 0 after normalization. The articles were classified into three sections: 'negative' for articles with

Figure 2 Proportion of sentiment categories (own illustration)

a sentiment score below -0.3, 'neutral' for sentiment scores between negative and positive 0.3, and 'positive' for articles with sentiment scores above 0.3.

Figure 2 shows the proportion of each category. With this classification 26'623 articles or 35.4 % were classified as 'negative', 20'244 articles or 26.9 % were classified as 'neutral' and 28'318 articles or 37.7 % were classified as 'positive'.

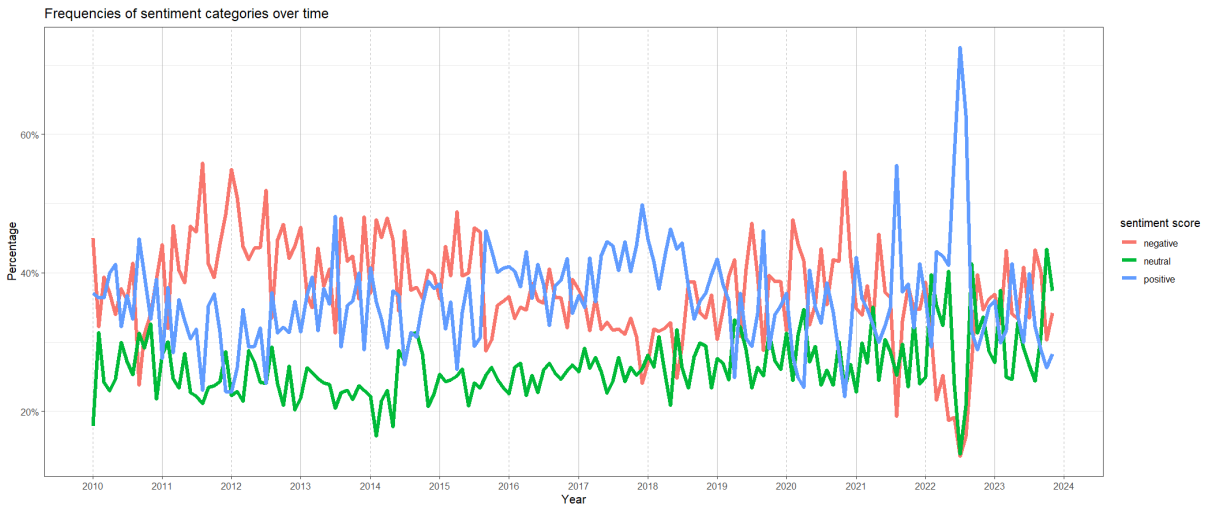

Figure 3 Line plot showing the change of proportions of the different sentiment categories over time (own illustration)

The proportion of sentiment categories in media articles about asylum migration changes over time due to various factors. Figures 3 and 4 display this change in proportion, with the x-axis representing time in years and the y-axis representing monthly percentages of the different categories. The two plots differ in their presentation of the data, with figure 3 being a simple line plot and figure 4 being a stacked area plot.

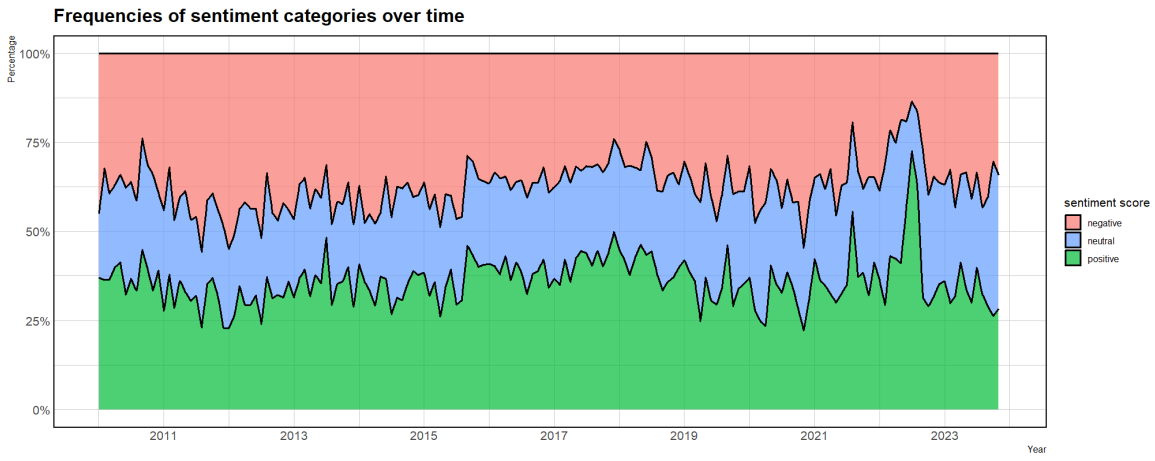

Figure 4 Stacked area plot showing the change of proportion of the different sentiment categories over time (own illustration)

It is difficult to recognize a trend, as the proportions of the three categories fluctuate strongly. However, there is a clear spike in positive articles from the beginning to mid-2022. This is probably linked to the outbreak of war in the Ukraine and the big wave of solidarity with Ukraine in Switzerland. A similar phenomenon but significantly weaker can be observed in mid-2015, during the peak of the refugee crisis in Europe and Switzerland.

To further analyze the trend in the swiss media coverage of asylum migration, we calculated an average sentiment score over the period from 2010 to 2023 which can bee seen in figure 5. The sentiment score, ranging from -0.9 to 0.3, is shown on the y-axis. The blue line indicates the trend or average sentiment score over time, it indicates a slightly positive tonality development, albeit at a low level. This suggests that the coverage of the migration issue in Switzerland has been predominantly negative over the years. From 2013 to 2017, the average sentiment score increased from -0.3 to around -0.075, whereas from 2017 to 2021, it remained relatively stable at around -0.075. Between 2021 and 2023, the average sentiment score shifted into the positive range, indicating a more positive tonality.

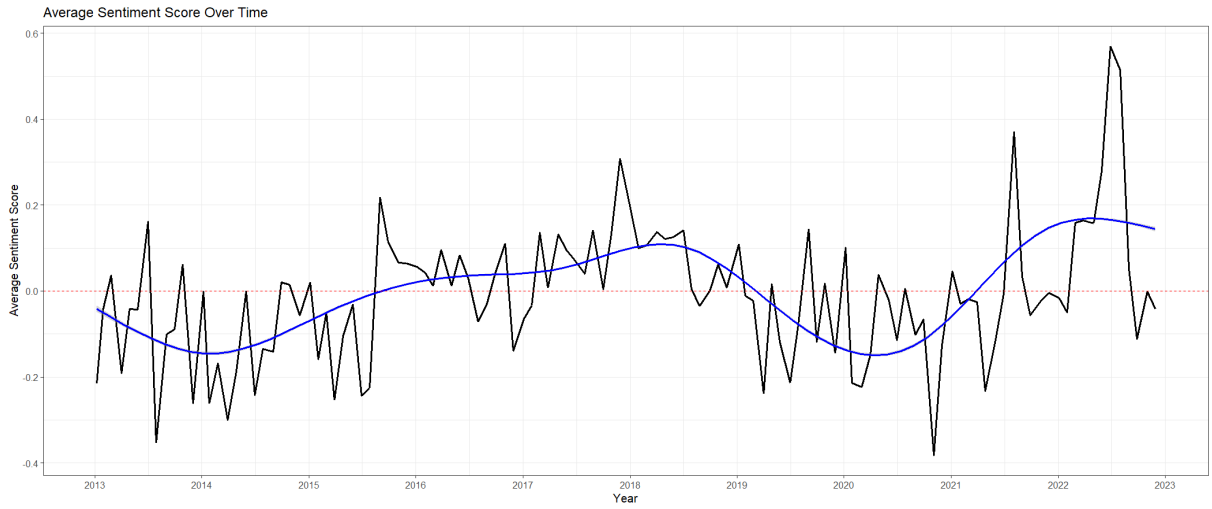

Figure 5 Progression of the average sentiment score (own illustration)

Correlation to asylum statistics

Figure 6 displays the correlation between the average sentiment score and the absolute number of asylum immigrants over time. The x-axis indicates the period of analysis from 2013 to 2022. The red trend line represents the average sentiment score, while the blue trend line represents the absolute immigration figures. Both trend lines were z-standardized and scaled, as shown on the y-axis. The trend lines follow somewhat a similar pattern, starting in negative territory and rising until 2017, followed by a negative development until mid-2020, after which they move in a positive direction again.

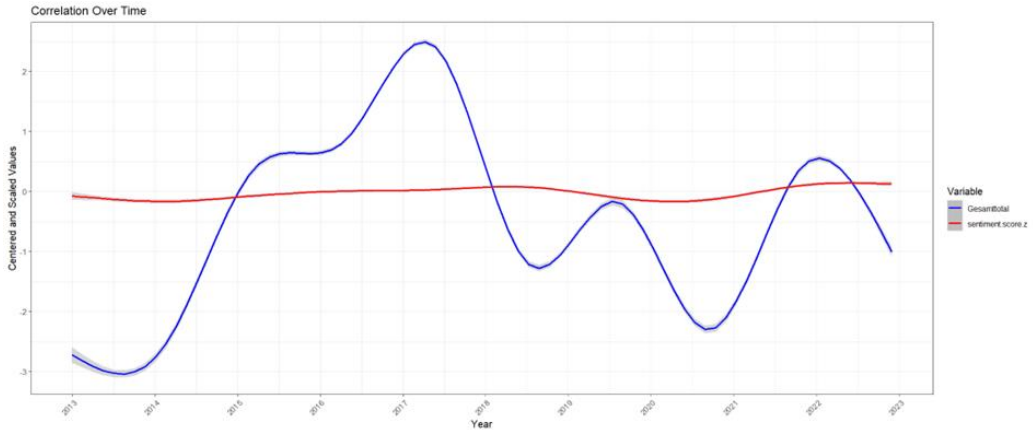

Figure 6 Correlation between total amount of asylum immigrants and sentiment score (own illustration)

But in summary, there is no statistically significant correlation between the absolute number of immigrants and the average sentiment score, as the variances overlap and the mean values intersect at two points.

Regression

The initial linear regression (see table 1) analyzes the significance of the sentiment score. It regresses the total number of asylum applications for each month between 2013 and 2023 with the z-standardized sentiment score. No significant results were found.

Linear Regression Results

	Dependent variable:
	sentiment.score.z
Gesamttotal	0.00004
	(0.00003)
Constant	0.014
	(0.010)
Observations	62,042
R^2	0.00003
Adjusted R^2	0.00001
Residual Std. Error	0.996 (df = 62040)
F Statistic	1.578 (df = 1; 62040)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1 Linear regression results on the sentiment score and the total of asylum seekers (own calculation)

The second linear regression analyses (see table 2) the connection between the sentiment score and the origin of the immigrants mentioned in the articles. The sentiment scores for immigrants from Africa, America, Asia, Europe, and Oceania were analyzed. The regression model explains only 0.2 % of the variation in the sentiment score, which is influenced by the origin of the immigrants from the respective continents. Statistically significant results were found for America, Europe and Oceania, but not for Africa and Asia.

A sentiment score of 0.002 was determined for European immigrants. This estimate has a confidence level of 99 %. Similarly, immigrants from America have a sentiment score of 0.009 and are expected to have a positive tonality with a confidence level of 99 %. Immigrants from Oceania are expected to get a more positive tonality, with a score of 0.071 in the model. Altough not on the same confidence level as the other two continents. Africa and Asia do not seem to have any impact.

To summarize, newspaper reports tend to have a positive tone when mentioning immigrants from America, Oceania and Europe. As for immigrants from Africa and Asia, no statement can be made due to the lack of statistical significance.

Linear Regression Results

	Dependent variable:
	sentiment.score.z
Afrika	0.00003
	(0.0001)
Amerika	0.009^{***}
	(0.001)
Asien	0.00005
	(0.00004)
Europa	0.001^{***}
	(0.0002)
Ozeanien	0.071^{*}
	(0.040)
Constant	-0.026^{**}
	(0.013)
Observations	62,042
R^2	0.003
Adjusted R^2	0.002
Residual Std. Error	0.994 (df = 62036)
F Statistic	31.802^{***} (df = 5; 62036)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2 Linear regression results on the sentiment score and the continent of origin (own calculation)

Discussion and Conclusion

The goal of this project was to analyze the Swiss media coverage of the refugee issue in the past decade. Using the database Swissdox and their API connection we gathered a large dataset with articles containing certain keywords. To filter this data, we used a LDA model to assign each article a topic and only kept those with a relevant topic. Based on the SentiMerge dictionary we conducted a sentiment analysis of the filtered dataset. We could show that major events such as the outbreak of the Ukraine war in 2022 or the refugee crisis in 2015 had a significant effect on the tonality of media articles. Furthermore, we found a correlation of asylum immigration numbers and the tonality of the media covering refugees and differences in tonality depending on the origin continent of the asylum migrants. However, the effects are incredibly small, and we are aware that the methodology of this paper is not beyond all doubt. Further, more detailed research is needed to understand all the processes and relationships influencing each other. Nevertheless, we think, that our explorative approach isn't worthless and can be used as a foundation for further research.

Bibliography

BfS, Bundesamt für Statistik. 2023. “Printmedien: Auflage- und Leserzahlen ausgewählter Tages- und Sonntagszeitungen, nach Sprachregionen - 2014-2022 | Tabelle.” *Bundesamt für Statistik*. https://www.bfs.admin.ch/asset/de/24785125 (January 9, 2024).

Emerson, Guy, and Thierry Declerck. 2014. “SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework.” In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 30–38. https://doi.org/10.3115/v1/W14-5805 (January 9, 2024).

SEM, Staatssekretariat für Migration. 2023. “Asylstatistik.” *Statistiken*. https://www.sem.admin.ch/sem/de/home/publiservice/statistik/asylstatistik.html (January 9, 2024).

SRF, Schweizer Radio und Fernsehen. 2015. “Wahlen 15 - SVP warnt vor Flüchtlingsansturm.” *Schweizer Radio und Fernsehen (SRF)*. https://www.srf.ch/news/wahlen-15-svp-warnt-vor-fluechtlingsansturm (January 9, 2024).

Swissdox AG. “Datenanalyse mit Swissdox@LiRI.” *Swissdox*. https://swissdox.ch/datenanalyse-mit-swissdoxliri/ (January 9, 2024).

Statement of authorship

„We hereby declare that each of us has contributed their part to this thesis without any help from others. Furthermore, we declare that we have written it without the use of aids other than those stated above. We have mentioned all used sources and cited them correctly according to the established academic citation rules. We are aware that otherwise, according to the University Act, the Senat is entitled to revoke the degree awarded on the basis of this thesis.“

Bern, 1/9/2024 Simon Bernhard

Bern, 1/9/2024 Kevin Jan Schläpfer