AntiSemRO: Studying the Romanian expression of Antisemitism

Anonymous ACL submission

Abstract

With far-right ideology rising in popularity, online environment embodies hateful attitudes. The Covid-19 pandemic and the violent wars in Ukraine and Palestine contributed to a growth in antisemitic discourse. This study introduces an annotated dataset for the study of antisemitic hate speech in Romanian along with several baseline models using classical machine learning models and transformer models for the classification of antisemitic discourse in the online medium.

1 Introduction

Discourse does not exist in a vacuum and can be as important as action. It creates and propagates ideas. Unfortunately, some ideas derive from places of hate and make life difficult for certain people. The ones in power positions have the capacity to develop a framework to protect vulnerable individuals who are the target of such discourse. However, often enough, those in power who are able to balance these situations, do the opposite.

As per the handbook comprised by the Council of Europe (Pausch et al., 2022), there is a growing anti-democratic tendency worldwide. This phenomenon is also enhanced through the hateful discourse often practiced by persons who have right-wing values as per this study by (Knüpfer, 2024) . Unfortunately, right-wing ideology flourishes in unstable, unequal, poverty-stricken societies the studies by (Jay et al., 2019) and (Franc and Pavlović, 2023) shows. The Worldbank poverty and equity brief [1] from April 2023 reports that the rate of Romanians who live at risk-of-poverty is 22.6%. This rate is the highest in the EU. Another aspect is the inequality driven by the inability of the state to raise the quality of life in the rural area where 75% of the poorest live.

There has been a steady interest for hate/offensive/toxic speech detection in the NLP academic environment ((Schmidt and Wiegand, 2017; Jahan and Oussalah, 2023). However, there is a lack of research on Romanian language in the area of antisemitic discourse production. In our case, limited research can be easily motivated by the scarcity of resources. To our knowledge, few datasets are available. Annotated data is even more rare. We aim to provide a novel dataset for the study of antisemitic speech and a baseline for text classification. We will train several machine learning models both traditional and Transformers-based.

2 Related Work

This paper proposes a novel annotated dataset for the study of antisemitism based on the particular outlook of how this phenomenon has been manifesting in Romania along the years. Major events and certain periods in the history of Romania played an important role in the ways in which online users employ of an array of tropes to paint the actions and the identity of Jewish people. We realise that these tropes differ from culture to culture. The motivation section contains several events and persons important for the development of antisemitic discourse.

Tripodi et al. (2019) dive into an incursion on French periodicals and books in order to retrieve the biases in the texts of the 18-20th centuries. They performed embedding projections over 6 categories. The categories are related to the domains in which antisemitic bias often appears: religious, economic, socio-political, racial, conspiratorial and ethic.

Riedl et al. (2022) built the case for how social media platforms offer antisemitism "affordances" in the shape of platform-specific functionalities. They used Twitter for their study and showed how hashtags, re-tweets and quote-tweets each help to

WARNING: This paper contains discriminatory language.

[1] https://databankfiles.worldbank.org.

the propagation of particular types of antisemitic discourse.

Steffen et al. (2022) published a German dataset for automated detection of antisemitic and conspiracy-theory content. Their work developed an annotation scheme for their dataset and pointed out important definitions for the underlying concepts related to antisemitic discourse.

Chandra et al. (2021) also collected two datasets from Gab and Twitter in order to train a multimodal deep learning model based on the categories proposed by Brustein (2003), namely: political, economic, religious and racial.

3 Motivation

The European Union Agency for Fundamental Rights reports that there is a lack of systematic data collection on antisemitism (for Fundamental Rights, 2023). Romania in its National strategy for preventing and combating antisemitism, xenophobia, radicalisation and hate speech 2021-2023 discusses an action plan to mitigate this problem. However, it is centered on manual intervention and monitoring [2]. Therefore, we wish to see whether we are able to provide an automatic method to detect antisemitic discourse and both a quantitative and qualitative overview of this type of discourse.

First, Romania has a far-right past with the Iron Guard movement that performed its activity during the 1930s. The most influential personality of this movement is, without doubt, Corneliu Zelea Codreanu, who is still present in the public discourse. Another figure important for the right-wing movement is Ion Antonescu, Prime-Minister and Ruler during most of WW2, who is responsible for the Holocaust in Romania. After the WW2 until the end of the 1980s the communist dictatorship left the country in shambles. This created an unbalanced environment where new political parties struggled for power. Social policies were barely put in place to cover for the poverty in which people were unable to make ends meet.

Poverty and education are highly correlated ((Mihai et al., 2015)) and the lack of it can make people vulnerable to prejudices ((Wodtke, 2012)). Education plays a crucial role in tackling this kind of attitudes. Romania has been in a ceaseless restructuring of the education system after the fall of the communist regime. However, the government's public expenditure on education is still lower than the EU average. 40% of Romanian students are functionally illiterate and there is a proven correlation between illiteracy and poverty (Thengal, 2013; Lal, 2015). So, poverty and lack of education are both great issues in Romania that contribute to the rise of extremist attitudes.

The last European elections show that the Romanian far-right party has been growing in popularity [3]. Therefore, we wish to start looking into the phenomenon of antisemitic discourse by publishing a dataset and training several text classification models for antisemitism detection.

4 AntiSemRO Corpus

The dataset presented in this study will be made available on Github. The 2162 posts were obtained using Crowdtangle from popular Romanian Facebook groups. We filtered the dataset by a list of keywords:evreu(*jew*), evrei(*jews*), evreul(*the jew*), evreii(*the jews*), evreilor(*jews'*), ovreu (archaic term for *jew*), ovrei(archaic term for *jews*), jidan(pejorative for *jew*), jidanul, jidanului, jidani, jidanii, jidanilor, jidanca, jidance, jidancelor, sionisti, sionism, zionism, chazar (person from a Turkic tribe who are mostly Jews), chazari, kazar, khazari, iudeo-masonic, iudeo-masonica, Holocaust, Holocaustului, Holocaustul, Holocau, Pogrom(relentless attacks organised by a mass of a militia or an organization against a minority) Pogromul, Pogromuri, Pogromului, Pogromurile, iudeu, jidov, semit, kipa, kipah, chipa, legionar, TLC[4], Traiasca Legiunea si Capitanul(Long Live the Legion and the Captain - slogan of the Iron Guard), Trăiască Legiunea și Căpitanul, CZC, Corneliu Zelea Codreanu. The different versions of the dataset are available on Github[5]

5 Annotation scheme

The annotations were done by two researchers from the "Elie Wiesel" National Institute for the Study of the Holocaust in Romania. They have a background in Sociology and Political Science hence they are able to pick the most subtle forms of hate speech and finely label the posts. Based on the studies by (Tripodi et al., 2019) and (Shafir, 2002)

[2] https://www.gov.ro/fisiere/programe_fisiere/Raport_final_strategie_mai_2022.pdf

[3] https://www.politico.eu/europe-poll-of-polls/romania/

[4] Traiasca Legiunea si Capitanul

[5] https://github.com/tobecompleted

	Label	%	No. of occurrences	Mean of words per category
Neutral: Unrelated	786	36.29%	786	326.34
Neutral or Informative	568	26.22%	568	220.65
Neutral: Ethnic Humour	47	2.17%	50	152.23
Neutral: Ambiguous	87	4.02%	87	281.48
Positive: Historical Awareness	341	15.74%	344	161.61
Negative: Holocaust: Minimization and trivialization of the Holocaust	91	4.20%	96	**619.26**
Positive: Confessions and solidarity	57	2.63%	58	154.65
Positive: Pro-Israel/Sionist political activism	52	2.40%	53	152.37
Negative: Political/economic antisemitism	49	2.26%	61	218.29
Negative: Reframing Nazism/fascism/legionarism	26	1.20%	44	332.58
Negative: Religious antisemitism	24	1.11%	38	281.08
Negative representation of Jewish people	19	0.88%	39	171.84
Negative: Judeo-Bolshevism	9	0.42%	17	266.89
Positive extremist: Extreme Pro-Israel	6	0.28%	6	297.67

Table 1: Frequency of texts per each category and mean of words per category. The table is valid for the pre-processed dataset without duplicates.

we devise our own annotation scheme which has some other categories not present in other studies. Some of the questions we ask when looking at the data are: how does the it portray the Jewish community and the history of the Holocaust? does it incite to hatred and violence? does it perpetuate negative antisemitic stereotypes? does it negate or trivialise the Holocaust and the suffering of the Jews? These are all the labels contained in the dataset: Neutral: Unrelated;Neutral or Informative ;Neutral: Ethnic Humour Neutral: Ambiguous; Positive: Historical Awareness; Negative: Holocaust: Minimization and trivialization of the Holocaust; Positive: Confessions and solidarity; Positive: Pro-Israel/Sionist political activism; Negative: Political/economic antisemitism; Negative: Reframing Nazism/fascism/legionarism; Negative: Religious antisemitism; Negative representation of Jewish people; Negative: Judeo-Bolshevism; Positive extremist: Extreme Pro-Israel. However, for the text classification task we use three big classes: Neutral, Negative and Positive. We do this because at the moment we do not a balanced amount posts. We also want to underline the difficulty to annotate antisemitic content. The censorship put in place by social media platforms pushes users to find subtler ways to express antisemitic prejudice. Therefore, annotating, detecting and truly understanding this type of manifestation takes special scrutiny.

6 Baseline Methods

We propose several baseline methods for the multilabel classification of antisemitic language using the new corpus we developed. To do this, we will use several encoding techniques, namely, bag-of-words, TF-IDF and BERT-based encoding alongside traditional machine learning techniques as Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Logistic Regression, Linear Support Vector Classification, K-Nearest Neighbours with Uniform Weight, and K-Neighbours with distant Neighbours.

7 Text Representations

We pair the traditional machine learning algorithms with Bag-of-Words (BOW) and Term Frequency–Inverse Document Frequency (TF-IDF). These encoding methods are language independent and help us model antisemitism better by keywords. For the Transformers models we use BERT text representations. The two available options for Romanian language are Multilingual BERT and Romanian BERT. Multilingual BERT developed by (Devlin et al., 2019) provides complex representations of texts containing information about context, syntax, and semantics. This kind of text representation performs well for low-resource languages like Romanian and they are widely used for text classification. Multilingual BERT was trained using Wikipedia data in 102 languages. Romanian BERT has been introduced by (Dumitrescu et al., 2020). This model is trained on a larger Romanian corpus and its tokenizer is better for handling Romanian due to using fewer tokens.

8 Experiments

As we struggle with both the size of our dataset and the percentage of actual antisemitic content we identified in the data we labelled, we apply a truncation method on our data as a data augmentation procedure as per (Sun et al., 2020). The dataset is split into training data and testing data. The training set contains 2958 neutral samples, 703 positive and 188 negative. The test set contains 328 neutral samples, 78 positive and 21 negative. As we have

Table 2: Results for antisemitic language detection on AntisemRO. We report Precision, Recall and F_1 for each model on the three classes (Neutral, Positive and Negative) and Macro F1-Score.

	Neutral			**Positive**			**Negative**			
Model	Precision	Recall	F1	Precision	Recall	F_1	Precision	Recall	F_1	Macro-F_1
BOW + Bernoulli NB	0.83	0.98	0.89	0.72	0.29	0.41	0.00	0.00	0.00	0.81
BOW + Multinomial NB	0.80	1.00	0.89	1.00	0.06	0.12	0.00	0.00	0.00	0.79
BOW + Linear SVC	0.89	0.99	0.94	0.92	0.62	0.74	1.00	0.32	0.48	0.89
BOW + k-NN w/uniform weight	0.89	0.97	0.93	0.82	0.65	0.72	1.00	0.32	0.48	0.89
BOW + k-NN w/distant neigh	0.84	0.87	0.85	0.35	0.35	0.35	0.43	0.14	0.21	0.75
Fine-tuned M-BERT	0.74	0.84	0.79	0.81	0.71	0.76	0.73	0.60	0.63	0.80
Fine-tuned Ro-BERT	0.71	0.69	0.70	0.91	0.92	0.92	0.63	0.57	0.60	0.78

already mentioned, our dataset is quite small, therefore we need to perform a 5-fold cross-validation for each model. The BERT models are used together with the AdamW optimizer and a 0.00001 learning rate with 50 warm-up steps. We train each model for 5 epoques. The table above shows the performance of each model for the 5 splits for Precision, Recall, F1-Score and Macro F1. Out of the traditional machine learning models the best performance is recorded for the KNeighbors with Distant Neighbors. Models perform best identifying neutral comments which is to be expected due to a bigger number of neutral samples.

9 Results and Limitations

We are aware that at the moment we are limited by the quantity of our data and the inability to annotate more due to time constraints. The current results are heavily influenced by how imbalanced our dataset is. We believe that the results obtained using the two BERT models are the most reliable as we have uniform values for precision and recall across all classes.

10 Conclusion

The process of collecting and annotating this dataset proves that there is plenty to discover about the phenomenon of antisemitic discourse. There will be further research into how the different types of antisemitic speech are expressed, their frequency and what their particularities are. At the basis of our study is the desire to be able to quantify these expressions and form a reliable opinion on this subject. After having these answers it will be possible to inform competent institutions and create a robust plan for tackling antisemitic attitudes.

References

William I. Brustein. 2003. *Roots of Hate: Anti-Semitism in Europe before the Holocaust*. Cambridge University Press.

Mohit Chandra, Dheeraj Reddy Pailla, Himanshu Bhatia, AadilMehdi J. Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. *CoRR*, abs/2104.05947.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

European Union. Agency for Fundamental Rights. 2023. *Antisemitism: Overview of Antisemitic Incidents Recorded in the European Union 2012-2022 : Annual Update*. Publications Office of the European Union.

Renata Franc and Tomislav Pavlović. 2023. Inequality and radicalisation: Systematic review of quantitative studies. *Terrorism and Political Violence*, 35(4):785–810.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Sarah Jay, Anatolia Batruch, Jolanda Jetten, Craig McGarty, and Orla T. Muldoon. 2019. Economic inequality and the rise of far-right populism: A social psychological analysis. *Journal of Community & Applied Social Psychology*, 29(5):418–428.

Curd Benjamin Knüpfer. 2024. Hate Speech and Political Violence: Far-Right Rhetoric from the Tea Party to the Insurrection by Brigitte L. Nacos, Yaeli Bloch-Elkon and Robert Y. Shapiro. *Political Science Quarterly*, 139(2):298–299.

B Suresh Lal. 2015. The economic and social cost of illiteracy: an overview. *International Journal of Advance Research and Innovative Ideas in Education*, 1(5):663–670.

Mihaela Mihai, Emilia Ţiţan, and Daniela Manea. 2015. Education and poverty. *Procedia Economics and Finance*, 32:855–860. Emerging Markets Queries in Finance and Business 2014, EMQFB 2014, 24-25 October 2014, Bucharest, Romania.

Markus Pausch, Patricia Hladschik, Rasha Nagem, and Filip Pazderski. 2022. Resilience against antidemocratic tendencies through education handbook for youth and social workers.

Martin J. Riedl, Katie Joseff, Stu Soorholtz, and Samuel Woolley. 2022. Platformed antisemitism on twitter: Anti-jewish rhetoric in political discourse surrounding the 2018 us midterm election. *New Media & Society*, 0(0):14614448221082122.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

M. Shafir. 2002. *Between Denial and "comparative Trivialization": Holocaust Negationism in Post-communist East Central Europe*. Analysis of current trends in antisemitism. Hebrew University of Jerusalem, Vidal Sassoon International Center for the Study of Antisemitism.

Elisabeth Steffen, Helena Mihaljević, Milena Pustet, Nyco Bischoff, María do Mar Castro Varela, Yener Bayramoğlu, and Bahar Oghalai. 2022. Codes, patterns and shapes of contemporary online antisemitism and conspiracy narratives – an annotation guide and labeled german-language dataset in the context of covid-19.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Niranjan Thengal. 2013. Social and economic consequences of illiteracy. *International Journal of Behavioral Social and Movement Sciences*, 2(2):124–132.

Rocco Tripodi, Massimo Warglien, Simon Levis Sullam, and Deborah Paci. 2019. Tracing antisemitic language through diachronic embedding projections: France 1789-1914. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 115–125, Florence, Italy. Association for Computational Linguistics.

Geoffrey T. Wodtke. 2012. The impact of education on intergroup attitudes: A multiracial analysis. *Social Psychology Quarterly*, 75(1):80–106.