AntiSemRO: Studying the Romanian Expression of Antisemitism

Anca Dinu
Faculty of Foreign Languages and Literatures,
University of Bucharest,
Romania
ancaddinu@gmail.com

Andreea-Codrina Moldovan
Interdisciplinary School of Doctoral Studies,
University of Bucharest,
Romania
moldovanandreeacodrina@gmail.com

Adina Marincea
"Elie Wiesel" National Institute for the Study of the Holocaust in Romania,
Romania
adina.marincea@gmail.com

Abstract

This study introduces an annotated dataset for the study of antisemitic hate speech and attitudes towards Jewish people in Romanian, collected from social media. We performed two types of annotation: with three simple tags ('Neutral', 'Positive', 'Negative'), and with five more refined tags (Neutral', 'Ambiguous', 'Jewish Community', Solidarity', 'Zionism', 'Antisemitism'). We perform several experiments on this dataset: clusterization, automatic classification, using classical machine learning models and transformer-based models, and sentiment analysis. The three classes clusterization produced well grouped clusters, while, as expected, the five classes clusterization produced moderately overlapping groups, except for 'Antisemitism', which is well away from the other four groups. We obtained a good F1-Score of 0.78 in the three classes classification task with Romanian BERT model and a moderate F1-score of 0.62 for the five classes classification task with a SVM model. The lowest negative sentiment was contained in the 'Neuter' class, while the highest was in 'Zionism', and not in 'Antisemitism', as expected. Also, the same 'Zionism' category displays the highest level of positive sentiment.

1 Introduction

There has been a steady interest in the detection of hate / offensive / toxic speech (Schmidt and Wiegand, 2017; Jahan and Oussalah, 2023) in the last decade. The automatic detection of these types of discourse, especially social media content including racist, homophobic, sexist, and extremist speeches, has seen great progress recently, mostly in major languages such as English, German, Spanish, or Arabic. However, for less resourced languages, like Romanian, there is still a need to develop databases that can be used for automatic detection of such discourse, in an effort to study and prevent it. This is an important direction in Natural Language Processing, since discourse does not exist in a vacuum and can be as important as action. It creates and propagates ideas, and if these ideas derive from places of hate, then it can make life difficult for certain groups of people.

This article attempts to fill a part of that need, in particular for the area of Romanian antisemitic discourse identification. We construct a novel dataset for the study of antisemitic speech and also about attitudes towards the Jewish community, manually annotate it with relevant tags, and then perform several tasks on it: clustering, sentiment analysis, and text classification with both traditional machine learning techniques and deep learning transformer-based models.

2 Related Work

Similar studies, in the NLP domain, have been carried out on major languages such as English, French, or German, but, to our best knowledge, none on Romanian.

Riedl et al. (2022) built the case for how social media platforms enable the spread of antisemitism in the shape of platform-specific functionalities. They used Twitter as data source, and showed how hashtags, re-tweets, and quote-tweets each help to the propagation of particular types of 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 posts.

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 on six categories related to the domains in which antisemitic bias often appears: reli-

WARNING: This paper contains discriminatory language.

Proceedings of Recent Advances in Natural Language Processing,pages 291–298
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-036

gious, economic, sociopolitical, racial, conspiratorial, and ethic.

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

3 Motivation

The European Union Agency for Fundamental Rights reports that there is a lack of systematic data collection on antisemitism (FRA, 2023). Also, Romania, in its National strategy for preventing and combating antisemitism, xenophobia, radicalization and hate speech 2021-2023, discusses an action plan to mitigate this problem. However, it focuses on manual intervention and monitoring[1], lacking training data and automated methods. In particular, antisemitism in Romania is of interest, since the country has a far-right past with the Iron Guard movement that performed its activity during the 1930s. The most influential personalities of this movement were the leader of the Iron Guard, Corneliu Zelea Codreanu and Ion Antonescu, Prime-Minister and Ruler during most of the Second World War, who is responsible for the Holocaust in Romania. Both are still present in public discourse. Influential politicians appropriate their discourse for gaining political traction [2]. Moreover, the latest political events show that the Romanian far-right has grown in popularity[3], which may contribute to the rise of extremist attitudes. Therefore, this work aims to help prevent antisemitic discourse by building a Romanian antisemitism dataset and by training several text classification models for automatic antisemitism detection.

4 AntiSemRO Corpus

Table 1: Statistics for the 'Message' Column in AntiSemRo

Statistic	Value
Total Messages	2165
Average Length (chars)	1666
Median Length (chars)	959
Max Length (chars)	4911
Min Length (chars)	5
Average Length (words)	264
Median Length (words)	145
Max Length (words)	923

The corpus was gathered using Crowdtangle from popular Romanian Facebook groups, totaling 2165 posts. To obtain relevant posts, we filtered the initial crowd-sourced data by a manually produced list of keywords, containing the following words along with all their inflected forms (plurals, feminine forms, genitives, definite article forms, since in Romanian the definite articles are enclitic - they are attached at the end of the word, etc.): evreu / iudeu / semit (*jew*), ovreu (archaic term for *jew*), jidan / jidov (pejorative for *jew*), sionist (*Zionist*), sionism (*Zionism*), chazar / kazar (person from a Turkic tribe who are mostly Jews), iudeomasonic (*Judeo-Masonic*), Holocaust / Holocau (*Holocaust*), Pogrom (relentless attacks organised by a mass of a militia or an organization against a minority), kipa / kipah / chipa (*kipa*), legionar (member of the far-right organization Iron Guard, during the 1930s), Traiasca Legiunea si Capitanul(*Long Live the Legion and the Captain* - slogan of the Iron Guard), TLC (initials of the Iron Guard slogan), Corneliu Zelea Codreanu (name of the leader of the Iron Guard), CZC (name initials of Corneliu Zelea Codreanu).

The different versions of the dataset we called AntiSemRo are available on Github[4]. Table 1 shows some general statistics for the dataset AntiSemRo.

The metadata labels of the dataset are the following: 'URL', 'Message', 'Description', 'Label', 'Page Name', 'Page Created', 'Likes at Posting', 'Followers at Posting', 'Post Created', 'Post Created Date', 'Post Created Time', 'Type', 'Total Interactions', 'Likes', 'Comments', 'Shares', 'Love', 'Wow', 'Haha', 'Sad','Angry', 'Video Share Status', 'Is Video Owner?', 'Post Views', 'Total Views', 'Total Views For All Crossposts', 'Video Length', 'Total Interactions (weighted — Likes, Shares, Comments, Love, Wow, Haha, Sad, Angry, Care).

Based on the studies by (Tripodi et al., 2019) and (Shafir, 2002) we devise our own annotation scheme which has some other categories not present in other studies, which capture local history. 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.This was by no means an easy task. 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. The inter-annotator agreement, measured using Cohen's Kappa, is 0.67, which indicates a substantial level of agreement.

The annotation scheme included the following labels for the dataset:

- Neutral - 1491:
  - Unrelated - 786;
  - Informative - 568;
  - Ethnic Humour - 50;

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

[2] https://revdem.ceu.edu/2024/12/05/rise-of-calin-georgescu/

[3] https://www.bbc.com/news/articles/crk2xxzxkzxo

[4] https://github.com/grrrrah/AntiSemRO

  - Ambiguous - 87;

- Positive - 456:
  - Historical Awareness - 341;
  - Confessions and solidarity - 57;
  - Pro-Israel/Sionist political activism - 52;

- Negative - 218:
  - Minimization and trivialization of the Holocaust - 91;
  - Political/economic antisemitism - 49;
  - Reframing Nazism/Fascism/Legionarism - 26;
  - Religious antisemitism - 24;
  - Negative representation of Jewish people - 19;
  - Judeo-Bolshevism - 9.

As expected, these sub-labels were not well-balanced, since the 'Negative' posts are more infrequent than the 'Neutral' ones, which form the vast majority, and than the 'Positive' posts, which are approximately double in number than the 'Negative' posts.

In order to clarify the meaning behind these labels we offer a list of examples below.

- 1. Unproblematic or Ambiguous or Unclear

- a. Unrelated (does not provide information related to the research topic): - "Bible quotes where "Jews" is just the name of a section." - "Posts mentioning the legionaries without additional details." - "I found this small-sized, young male dog wandering among cars around 6:20 PM at the intersection of Avram Iancu and Deportation of Jews Street, near Balcescu Park."

- b. Informative (news or objective/neutral information): "The General Prosecutor's Office dismissed the case regarding the minimization of the Holocaust by AUR."

- c. Ambiguous or Unclear: "PLEASE, I BEG YOU! Listen to this scientist (born in Ukraine)... before this VIDEO gets deleted by Facebook. He is Jewish, one of TRUMP'S DOCTORS, banned from social media during the GLOBALIST PLANDEMIC..."

- 2. Ethnic Humor:

- a. Jokes about Jews: "A Jew moves to a small Catholic town. Every Friday, while Christians were fasting and eating only fish, the Jew was grilling steak after steak, driving them mad with the smell. Desperate, the Catholics decided to convert him, and after threats, pleas, and promises, they managed to convince him. They took him with great ceremony to the Church, where the priest sprinkled him three times with holy water, chanting, "Born a Jew, raised a Jew, now a Christian."The following Friday, while all the Catholics were fasting and eating only fish, a mouthwatering smell of grilled meat wafted from the converted Jew's house. Driven crazy, they went to the sinner's house to see how this was possible. There, they found the Jew in front of a big grill loaded with meat, sprinkling it energetically with water, chanting: "Born a cow, raised a cow, now fish."

- 3. Antisemitism

- a. Religious Antisemitism: "The arm of God mocked. People wouldn't have known or believed if someone told them that the poor, silent, despised prisoner was 'the arm of the Lord.' The Jews who persecuted Jesus often sang in their synagogues: 'You have a mighty arm. Strong is Your hand, and high is Your right hand' (Psalm 89:13)."

- b. Political/Economic Antisemitism: "Mutin, the Masonic Christian-Jew. The biggest GLOBALIST."

- c. "Judeo-Bolshevism": -"Jews are responsible for bringing communism to Romania/world." - "Here is the list of Jewish communists who led Romania during the years of the harshest dictatorship and repression! Why don't we have a trial for communism?"

- d. Generic Antisemitism (tendentious presentation of the Jewish community): "Source: Violence erupted on Sunday in Jerusalem between Orthodox Jews and Palestinian activists at the beginning of the 'Flag March,' also known as 'Jerusalem Day.' Over 70,000 nationalist Jews marched Sunday afternoon through and around the Old City of Jerusalem to mark Yom Yerushalayim (Jerusalem Day), some chanting racist slogans and clashing with Palestinians and police..."

- 4. Holocaust/Rehabilitation or Reinterpretation of Nazism, Fascism, Legionarism

- a. Minimization or Trivialization or Distortion or Denial of the Holocaust

- b. Reframing of Nazism or Fascism or Legionarism: Particularly in the context of the war in Ukraine or the COVID-19 pandemic: "PRO OR ANTI-NAZI? / It's almost comic if it weren't tragic. The neo-communists at the White House and Brussels support Zelensky's neo-Nazis. Joe Biden signed yesterday the law allowing the expedited shipment of military equipment to Ukraine. The law is similar to a program from World War II to help Europe resist Hitler."

- 5. Positive Image about Jews/Israel, Historical Awareness, and Activism

- a. Historical Awareness (Holocaust remembrance, affirmation, and awareness, personal histories, etc.): "AUSCHWITZ: The camp where NAZIS gassed 1.5 MILLION JEWS. It's good to remember the atrocities our ancestors went through. Never again. And here we have a godless individual like Putin making us relive the past. I hope he meets the same fate as Hitler!"
- b. Community Opinions / Community Solidarity / Criticism from Jews Towards Their Leaders: - "Naftali Bennett, ZERO hiding in the illegally built citadel, unaware that Jews are being attacked with stones and axes in the Jordan Valley." - "Yair Lapid threatens the Chabad people; the Hanukkahs at every street corner, crossroads, Hanukkah worldwide, and the Chabad houses everywhere, will be over. Your story about love for Israel is over. It seems illogical to threaten the most soulful Jews worldwide, who bring only honor and dignity to the country."

5 Experiments

For all the experiments, we extracted from the data two different sets.

For the first set, we used three main classes: 'Neutral', 'Negative' and 'Positive'. Since these three classes were not balanced, we randomly chose 150 posts per each class, 450 in total.

In an attempt to refine the three categories above, we extracted from the dataset five groups as following: 'Neutral', 'Ambiguous', 'Jewish Community Solidarity', 'Zionism' and 'Antisemitism'. We consider that there needs to be a distinction between neutral and ambiguous comments. 'Neutral' class includes two sub-labels, 'Unrelated' and 'Informative', while 'Ethnic humor' was incorporated into 'Ambiguous' class. The 'Positive' class has been split into two smaller categories that better encompass meaning. We have 'Historical awareness' and 'Confessions and Solidarity' combined into 'Jewish Community Solidarity'. Last but not least, the two labels indicating Zionist attitudes are now part of a distinct category, 'Zionism'. The 'Antisemitism' class incorporates all the labels from the 'Negative' class.

For each of the five categories, we tried to have as many texts as possible and still keep the data fairly balanced. In total, we have 674 samples, distributed as indicated in figure 1.

5.1 Clusterization

We first employed clusterization techniques in an attempt to evaluate if the data is suitable for automatic classification, that is, if there are relevant discriminating features in the selected categories.

We vectorized both datasets using TF-IDF and then we reduced the dimensionality with Principal Components Analysis (PCA).

As depicted in figure 2, the 'Positive', 'Neutral' and 'Negative' posts categories are well grouped, while, as

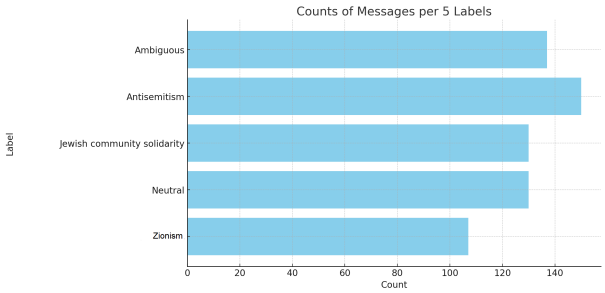

Figure 1: Number of posts in the 5 labels set

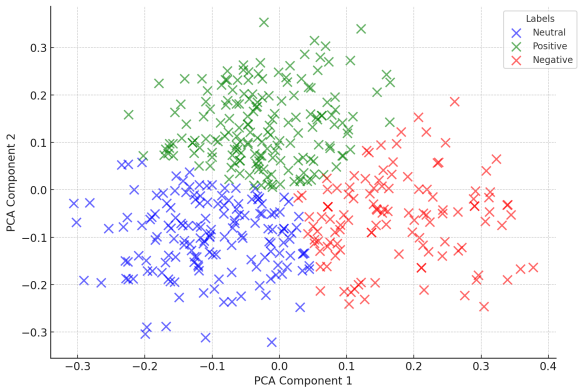

Figure 2: Clustering based on the dataset with 3 labels.

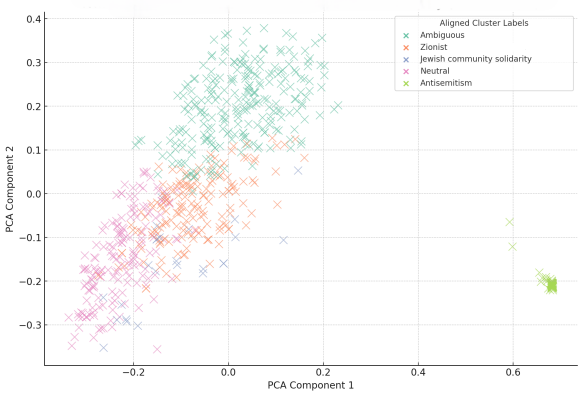

Figure 3: Clustering based on the dataset with 5 labels.

expected, the five categories from the second set are less distinguishable, as one can observe in figure 3. 'Antisemitism' is the most well defined category, far away from all the others, in the right lower corner of the image. 'Neutral' posts and posts regarding the Jewish community overlap, indicating a similar balanced tone of both categories. The 'Ambiguous' posts and the 'Zionist' post meet and partially overlap in the central area, suggesting the use of a similar, more radical content.

Figures 2 and 3 represent a good predictor for the performance of automatic classification task, which we detail in the next section.

5.2 Automatic classification

This section presents the experiments on the automatic classification of the 3-labels dataset ('Positive', 'Neutral', and 'Negative') and of the 5-labels dataset ('Neutral', 'Ambiguous', 'Jewish Community Solidarity', 'Zionism' and 'Antisemitism') from AntiSemRo.

5.2.1 Setting

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.

The dataset is split into training and testing sets using an 80-20 split from scikit-learn, ensuring that the same split can be reproduced using a random state parameter set to 42. We use a batch size of 16. The AdamW optimizer is employed with a learning rate (lr=2e-5) and epsilon (eps=1e-8). We use the typical BCE With Logits loss function and we train the model for 5 epochs, due to computational resource constraints.

5.2.2 Results

All the scores (precision, recall, F1, macro F1) for all the models we employed, are given in table 2 for the 3-label set and in table 3 for the 5-label set.

The performance across all models significantly decreases when moving from the 3-label set to the 5-label set, indicating that the models struggle more with the increased complexity and number of classes.

Both traditional text representation methods (TF-IDF and Word2Vec) perform comparably across different classifiers and SVM generally outperforms Logistic Regression and Random Forest.

For both datasets, with 3 and 5 labels, the transformer based model Romanian BERT performed better or equal than the multilingual transformer model and than the traditional machine-learning methods. In particular, the Romanian BERT excelled at classifying the 3-label set, particularly for the Positive and Negative labels, having the highest macro F1 score, of 0.78, at a great distance of the next best model, TF-IDF Random Forest, which scored only 0.54. For the 5-label task, Romanian BERT still scored the highest score, a modest 0.6 F1, at tie with the TF-IDF SVM. That suggests that the data base is not big enough for more than ternary classification.

5.3 Error analysis

n this subsection, we take a closer look at how the best-performing model, Romanian BERT, behaves and the misclassifications pattern it produces.

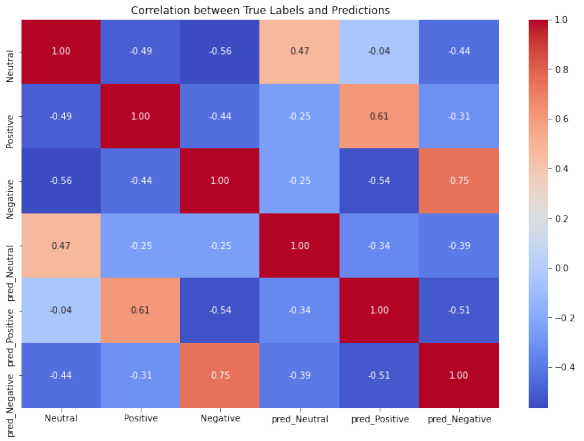

Figure 4: Heatmap showing the correlation between true labels and predicted labels for different categories for the 3-label dataset.

Figure 4 shows the correlations between the true labels and the predicted labels for all 3 categories. One can notice that the 'Negative' category is best predicted with a high correlation score of 0.75, followed by the 'Positive' category, with a lower score of 0.61. The 'Neutral' class is the most difficult to predict, with a correlation score of only 0.47.

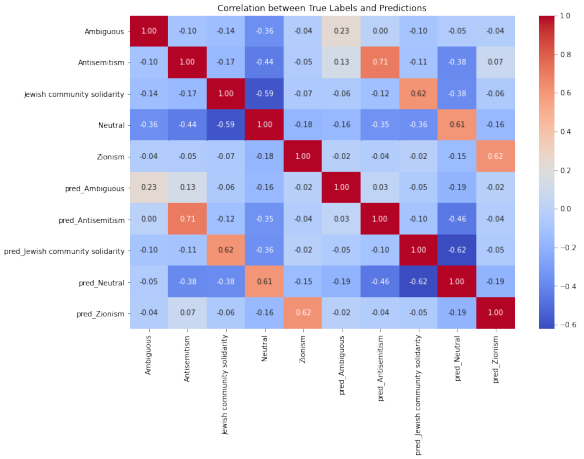

Figure 5: Heatmap showing the correlation between true labels and predicted labels for different categories for the 5 labels dataset.

In Figure 5 one can see the correlations between the true labels and the predicted labels for all 5 categories. The relationship between the Antisemitism class and its predictions indicate a positive correlation of 0.71, which is a decent score. Good results we see in identifying

Table 2: Model Performance for automatic classification of 3 label set (Neutral, Positive and Negative).

	Neutral			Positive			Negative			
Model	Precision	Recall	F1	Precision	Recall	F_1	Precision	Recall	F_1	Macro-F_1
TF-IDF - Logistic Regression	0.83	0.98	**0.90**	0.81	0.37	0.51	0.00	0.00	0.00	0.47
TF-IDF - Random Forest	0.84	0.97	**0.90**	0.76	0.41	0.54	0.80	0.11	0.20	0.54
TF-IDF - SVM	0.82	0.98	**0.90**	0.78	0.30	0.43	0.00	0.00	0.00	0.44
Word2Vec - Logistic Regression	0.80	0.96	0.87	0.41	0.13	0.20	0.00	0.00	0.00	0.36
Word2Vec - Random Forest	0.82	0.96	0.88	0.52	0.26	0.35	1.00	0.06	0.11	0.44
Word2Vec - SVM	0.79	0.99	0.88	0.38	0.02	0.04	0.00	0.00	0.00	0.31
BERT RO	0.71	0.69	0.70	0.91	**0.92**	0.92	0.63	0.57	**0.60**	**0.78**
BERT Multi	0.60	0.72	0.80	0.64	0.45	0.32	0.10	0.00	0.00	0.52

Table 3: Model Performance for automatic classification of 5 label set ('Neutral', 'Ambiguous', 'Jewish Community Solidarity', 'Zionism' and 'Antisemitism')

Model	**Neut**			**JCS**			**Anti**			**Zion**			**Ambig**			**M-F1**
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
TF-IDF - LR	0.41	0.68	0.51	0.59	0.55	0.57	0.57	0.65	0.61	0.65	0.32	0.43	0.63	0.63	0.63	0.55
TF-IDF - RF	0.42	0.73	0.53	0.8	0.55	0.65	0.57	0.55	0.56	0.44	0.32	0.37	0.55	0.63	0.59	0.54
TF-IDF - SVM	0.47	0.77	**0.59**	0.88	0.52	0.65	0.58	0.81	**0.68**	0.57	0.35	0.44	0.67	0.63	**0.65**	**0.62**
W2V - LR	0.38	0.64	0.47	0.54	0.52	0.53	0.49	0.58	0.53	0.35	0.24	0.28	0.12	0.05	0.07	0.38
W2V - RF	0.48	0.64	0.55	0.45	0.52	0.48	0.44	0.48	0.46	0.61	0.26	0.37	0.21	0.26	0.23	0.42
W2V - SVM	0.28	0.64	0.39	0.75	0.41	0.53	0.45	0.55	0.49	0.22	0.09	0.12	0.12	0.11	0.11	0.33
BERT RO	0.50	0.15	0.23	0.92	0.58	**0.71**	0.71	0.63	0.67	0.80	0.91	**0.85**	0.67	0.21	0.32	0.60
BERT Multi	1.00	0.06	0.11	0.95	0.55	0.70	0.79	0.59	**0.68**	0.80	0.91	**0.85**	1.00	0.21	0.35	0.54

'Jewish community solidarity' and 'Neutral' discourse. However, we notice that there is also some confusion between 'Jewish Community solidarity' and 'Neutral comments'.

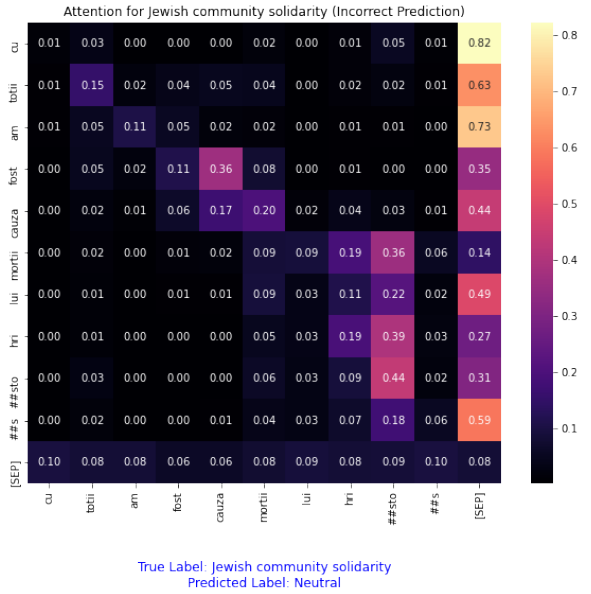

Figure 6: How attention is distributed in wrong predictions. The phrase translates as *We were all the cause of Christ's death.*

We give here two examples of incorrect prediction, typical for the misclassifications patterns, the first one which is semantically confusing, and the second one which is due to a tokenization error.

In Figure 6 we illustrate an example of the model wrongly predicting the label "Neutral" to a message which is actually tagged with "Jewish community solidarity". One can observed that the model paid considerable attention (gave consistent weights) to the words "cauza" (*the cause*), "mortii" (*death*), "hristos (tokenized hri-sto-s)" (*Christ*), words not associated with the true class "Jewish community solidarity", which partially explains the mistake.

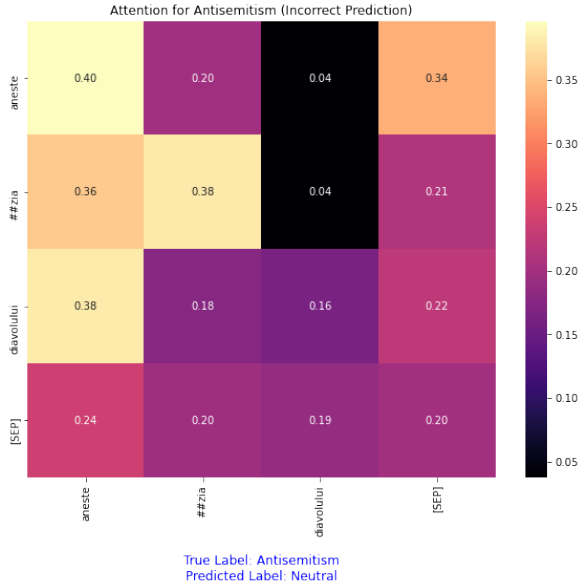

Figure 7: How attention is distributed in wrong predictions. The phrase translates as *The devil's anesthesia.*

Figure 7 illustrates an example of the model incorrectly predicting the 'Neutral' category for a text that actually pertains to 'Antisemitism'. In this case, the model heavily focused on 'aneste' and 'zia', wrongly tokenized as two separate words instead of just one, 'anestezie' (*anesthesia*), which lead to the incorrect pre-

diction.

5.4 Sentiment analysis

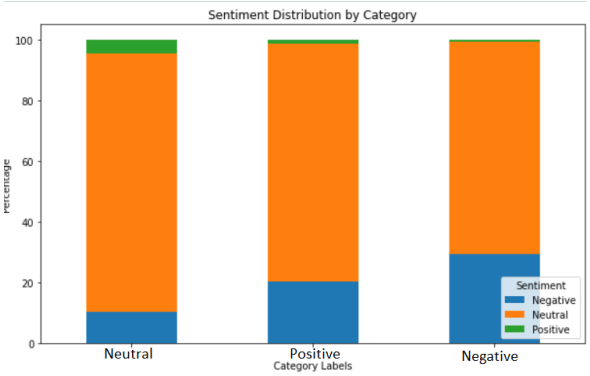

Figure 8: Summary of sentiment analysis for the 3-labels dataset.

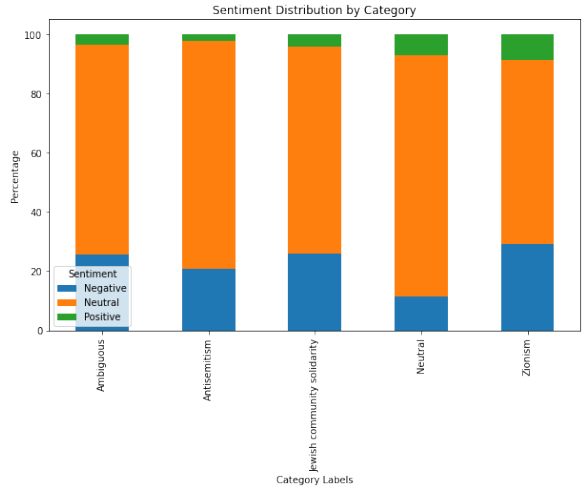

Figure 9: Summary of sentiment analysis for the 5-labels dataset.

We performed sentiment analysis with multilingual XLM-roBERTa-base model (Conneau et al., 2020) on both datasets, to analyze the predominant sentiment per label, leveraging the model's strong performance on multilingual NLP tasks, including sentiment classification.

As seen in Figure 8, the negative sentiment is, as expected, more prominent in the 'Negative' category, than in the 'Positive' one, and the lowest in the 'Neutral' class. The other way around happens with the positive sentiment, which has a remarkably low level. The most predominant sentiment in all three classes is, naturally, the neuter.

The sentiment analysis of the 5 labels set, depicted in figure 9, shows a more nuanced and surprising story. The lowest negative sentiment is contained in 'Neuter' class, while the highest is in 'Zionism', and not in 'Antisemitism', as expected. Also, the same 'Zionism' category displays the highest level of positive sentiment.

6 Limitations and Ethical Implications

One of the primary limitations of this study is the imbalance in the dataset, with the majority of posts falling into the 'Neutral' category, and significantly fewer posts categorized as 'Negative'. This imbalance can lead to a model bias where predicting less frequent categories is more difficult, potentially affecting the reliability of the model in real-world applications, if training is performed on the whole dataset.

The dataset was collected from a limited number of Romanian Facebook groups where subjects involving Jewish culture were often approached. While these sources are relevant, they may not capture the full spectrum of antisemitic discourse in Romanian online spaces. Other platforms, such as Twitter, Instagram, niche forums, were not included, therefore, generalizing the findings is not possible.

Another possible limitation of our study lies within the annotation area. While thorough, it was conducted by a small number of annotators with specific expertise. This could introduce subjective bias, particularly in interpreting subtle or ambiguous content.

We should also take into consideration model refinement. The BERT models, although generally effective, showed limitations in handling the multi-class categorization, particularly in distinguishing between closely related categories like 'Neutral' and 'Jewish Community Solidarity'.

It is crucial to consider the ethics of research involving such a sensitive topic. Even though CrowdTangle aggregates data from public sources, we must still consider the potential for privacy violations. Therefore, we ensure that the data is anonymized to protect the identities of the users. We also try to report our finding in ways that do not involve hate speech exemplification.

7 Conclusion

The process of collecting, annotating and analyzing this dataset proves that there is plenty to discover about the phenomenon of antisemitic discourse in online medium.

The AntiSemRo database is the first one of its type for Romanian and its analysis is promising, setting a baseline for further research into how the different types of antisemitic speech are expressed, their frequency and what their particularities are. Automatically identifying antisemitic discourse in a reliable manner would be a valuable tool for the competent institutions to create a robust plan for tackling antisemitic attitudes.

However, our findings indicate that while current models, particularly transformer-based models, show promise in identifying antisemitic content, challenges remain in distinguishing between nuanced categories and dealing with imbalanced data.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics.

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.

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

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

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.

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.