“Subverting the Jewtocracy”: Online Antisemitism Detection Using Multimodal Deep Learning

Mohit Chandra [1] Dheeraj Pailla [1] Himanshu Bhatia [1] Aadilmehdi Sanchawala [1]
Manish Gupta [1] Manish Shrivastava [1] Ponnurangam Kumaraguru [2]

[1] International Institute of Information Technology, Hyderabad, India
[2] Indraprastha Institute of Information Technology Delhi, India

DISCLAIMER

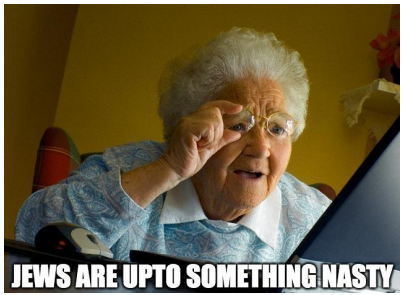
This presentation contains abusive/hateful content in the form of text and images, used only for illustrative purposes.
Viewer/reader discretion is advised.

INTRODUCTION

According to International Holocaust Remembrance Alliance (IHRA)

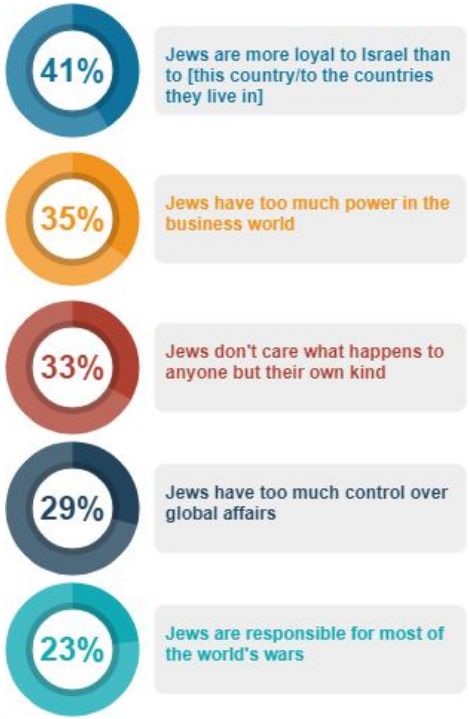
"Antisemitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, towards Jewish community institutions and religious facilities."

A post with benign text may as well be antisemitic due to a hateful image. Thus, it becomes essential to take a more holistic approach.



Text: Even grandma can see what's going on.

Text: I see the blews are at it again.



Source: ADL GLOBAL 100: AN INDEX OF ANTI-SEMITISM

CONTRIBUTIONS

We collect and label two datasets on online antisemitism gathered from Twitter and Gab with 3,102 and 3,509 posts respectively. Each post in both the datasets is labeled for presence/absence as well as antisemitism category.

We propose a novel multimodal system which learns a joint text+image representation and uses it for antisemitic content detection and categorization.

The presented multimodal system achieves an accuracy of **~91%** and **~71%** for the **binary antisemitic content detection task** on Gab and Twitter respectively. Further, for 4-class antisemitism category classification, our approach scores an accuracy of **~67%** and **~68%** for the two datasets respectively.

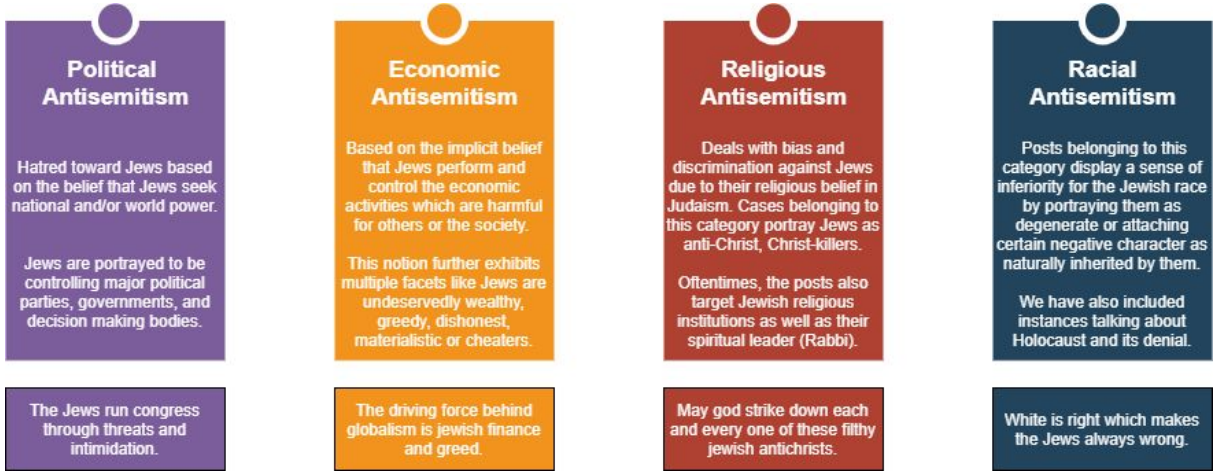
We provide a detailed qualitative study to analyse the limitations and challenges associated with this task and hate speech detection in general.

To the best of our knowledge, our work is the first in the direction of multimodal antisemitism detection.

ANTISEMITISM CATEGORIZATION

In addition to the binary classification of posts as antisemitic or not, we also classify the antisemitic posts in one of the four categories. We primarily referred to the categorization proposed by Brustein and augmented this categorization with additional inputs from the detailed IHRA's definition.

We categorize each antisemitic post in one of the four categories: 1) Political Antisemitism; 2) Economic Antisemitism; 3) Religious Antisemitism; 4) Racial Antisemitism



ONLINE ANTISEMITISM DATASETS

Data Collection

We choose Twitter and Gab to gather data for our study. We retained only those posts which contained text as well as images. Further, we ensured that each post included at least one term from a high precision lexicon.

Data Annotation

The examples were annotated on two levels after looking at the text as well as the image – (1) **binary label** (whether the example is antisemitic or not), and (2) **multiclass label** (if the example is antisemitic then assign the **respective antisemitism category**).

ONLINE ANTISEMITISM DATASETS

Data Statistics

As observed from Table 1, majority of posts in both datasets lie in either political antisemitism or the racial antisemitism category. We believe that this trend is inline with the phenomenon of 'New antisemitism'.

Overall, 84% of the total images had some form of text in them. This motivated us to use an OCR module in the proposed system. On average, post text has **~45** and **~27 words**, while the OCR output is **~50** and **~51 words** long, after pre-processing for Gab and Twitter respectively.

Table 1: Basic statistics for the two datasets.

	#Total Posts	#Antisemitic posts	#Political posts	#Economic posts	#Religious posts	#Racial posts
Twitter	3,102	1,428	639	183	124	482
Gab	3,509	1,877	736	118	144	879

Table 2: Frequent Unigrams and Bigrams for each of the Antisemitism Categories.

N-Grams	Political Antisemitism	Economic Antisemitism	Religious Antisemitism	Racial Antisemitism
Unigrams	jews, zionist, zog, israel, media, control. world, government, politics, conspiracy	jewish, money, cash, finance, wealth, business, bankers, kosher	jews, christ, jesus, killer, rabbi, expel, satan, christians, messiah	jews, jewish, fake, holocaust, hitler, white, hebrew, ridiculous, pinocchio
Bigrams	world domination, zionist jews, zionist occupied, terrorist zionist, jews state	jewish money, money politics, money everything, money launderers, zionist bankers	christ killer, read torah, jesus killer, ultra orthodox, rabbi israel, jewish ritual	jewish man, jews attacks, anti semitism, jewish people, race mixing

MULTIMODAL ANTISEMITISM CATEGORIZATION SYSTEM

Fig. 1 illustrates the architecture of our proposed multimodal system for online antisemitism detection with RoBERTa text+OCR encoder, ResNet-152 image encoder and the MFAS fusion module.

We experimented with BERT and RoBERTa for text encoding since they have been shown to lead to high accuracy across multiple NLP tasks. We also experimented with ResNet-152 and Densenet-161 for image encoding.

For getting the OCR output from the images we experimented with three different services(Google's Vision API, Microsoft's Computer Vision API and Open-source tesseract

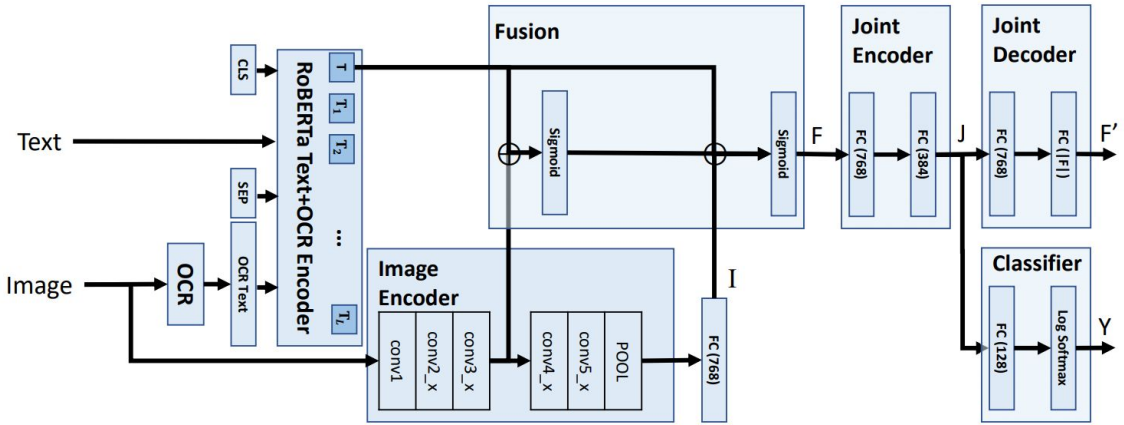


Figure 1: Proposed multimodal system architecture.

MULTIMODAL ANTISEMITISM CATEGORIZATION SYSTEM

Fusion

To combine the features obtained from the Text + OCR and the image encoder modules, T and I, we experiment with three different techniques of fusion – (1) Concatenation, (2) Gated MCB [Fukui et al., 2016] and (3) MFAS [Perez-Rua et al., 2019].

MFAS (Multimodal Fusion Architecture Search) first concatenates text and image representations from an intermediate hidden layer, applies a sigmoid non-linearity.

In the next step, it concatenates this with final layer text and image representations along with a sigmoid non-linearity.

Joint Encoder/Decoder and Classifier

The fused representation F is then passed through a series of Denselayers (768 and 384) to obtain a joint encoded vector J. J is fed to two modules: joint decoder and classifier.

The joint decoder aims to reconstruct F and uses MSE (mean squared error) loss, while the classifier aims to predict presence/absence of antisemitism or antisemitism category. We use the sum of these two losses to train the model.

EXPERIMENTS

Results using Text-only and Image-only Classifiers

We experimented with five popular pre-trained text embedding/network based classifiers and two pre-trained image network classifiers.

For the text-only classifiers we used GloVe [Pennington et al., 2014], FastText [Joulin et al., 2016], BERT [Devlin et al., 2018] and RoBERTa [Liu et al., 2019]. We also experiment with [Founta et al., 2019]'s method which is an attentional RNN model with GloVe embeddings.

For the image only classifiers we experimented with ResNet-152 [He et al., 2016]and DenseNet-161 [Huang et al., 2017].

Table 3: Comparison of (5-fold cross validation) performance of popular text-only and image-only classifiers. The best performing method is highlighted in bold separately for both the text and image blocks.

		Twitter				Gab			
		Binary		Multiclass		Binary		Multiclass	
		Accuracy	F-1	Accuracy	F-1	Accuracy	F-1	Accuracy	F-1
Text only	GloVe+Dense	.630±.009	.621±.013	.490±.013	.268±.025	.651±.027	.612±.040	.540±.031	.276±.018
	FastText+Dense	.540±.000	.351±.000	.467±.031	.223±.099	.566±.017	.429±.045	.532±.030	.269±.019
	GloVe+att-RNN	.583±.048	.552±.081	.416±.019	.239±.033	.630±.039	.624±.045	.460±.039	.240±.028
	BERT+Dense	.701±.015	.700±.016	**.669±.047**	**.676±.036**	**.889±.008**	**.889±.009**	.623±.025	.575±.038
	RoBERTa+Dense	**.733±.007**	**.733±.008**	.663±.039	.662±.050	.874±.010	.874±.010	**.632±.032**	**.583±.039**
Img only	ResNet-152	**.579±.014**	**.578±.015**	**.416±.028**	**.317±.040**	.587±.008	.583±.008	**.456±.020**	**.275±.010**
	Densenet-161	.567±.033	.566±.033	.405±.033	.281±.011	**.610±.017**	**.607±.015**	.446±.031	.274±.027

EXPERIMENTS

Results using Multimodal Classifiers

In this experiment we tested three different fusion mechanisms – (1) Concatenation, (2) Gated MCB [Fukui et al., 2016] and (3) MFAS [Perez-Rua et al., 2019] for our proposed architecture.

We also compared the performance of our proposed architecture with the baseline model from [Gomez et al., 2020] (FCM). FCM uses GloVe for text encoder and InceptionV3 for image encoder.

Table 4: Comparison of (5-fold cross validation) performance of multimodal classifiers with RoBERTa as text encoder and ResNet-152 as image encoder. We also compare the performance of the proposed architecture with a baseline from Gomez et al.[11] (FCM).

Method	Twitter				Gab			
	Binary		Multiclass		Binary		Multiclass	
	Accuracy	F-1	Accuracy	F-1	Accuracy	F-1	Accuracy	F-1
FCM [11]	.564±.015	.545±.038	.445±.006	.164±.022	0.607±0.014	.595±.028	.468±.005	.182±.027
Concatenation	.710±.012	.708±.013	.662±.027	.664±.027	.905±.005	.905±.005	.653±.052	.616±.046
Gated MCB	.690±.026	.683±.036	.679±.030	**.677±.043**	.904±.014	.903±.014	.654±.039	.618±.043
MFAS	**.715±.013**	**.714±.014**	**.680±.035**	.675±.023	**.906±.007**	**.906±.007**	**.665±.029**	**.625±.032**

QUALITATIVE ANALYSIS: ATTENTION VISUALIZATION

To gain better insights into the the proposed system, we visualized attention weights for both the Text + OCR using bertviz [Vig. 2019] and the Image encoder using GradCAM [Selvaraju et al., 2017].

We took an antisemitic example having the text content as **"some people have jew parasites embedded in their brains"** and the OCR text being **"liberals"**.

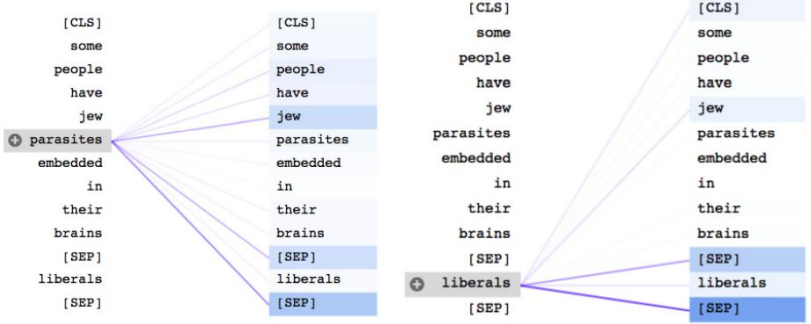


Figure 3: Text + OCR encoder module attention visualization

Figure 4: Image encoder module attention visualization (Best viewed in color)

QUALITATIVE ANALYSIS: ERROR ANALYSIS

Tables 6 and 7 show the confusion matrices for the proposed MFAS-based multimodal system for binary and multi-class cases respectively for Gab. Similarly, Tables 8 and 9 show confusion matrices for Twitter.

Table 6: Confusion matrix for the binary classification task (Gab). The entries represent the sum on test set examples over 5-fold cross validation.

		Predicted	
		Non-antisemitic	Antisemitic
Actual	Non-antisemitic	1470	162
	Antisemitic	167	1710

Table 7: Confusion matrix for the multiclass classification task (Gab). The entries represent the sum on test set examples over 5-fold cross validation.

		Predicted			
		Political	Economic	Religious	Racial
Actual	Political	441	49	33	213
	Economic	14	82	3	19
	Religious	9	1	102	32
	Racial	141	40	76	622

Table 8: Confusion matrix for the binary classification task (Twitter). The entries represent the sum on test set examples over 5-fold cross validation.

		Predicted	
		Non-antisemitic	Antisemitic
Actual	Non-antisemitic	1106	568
	Antisemitic	317	1111

Table 9: Confusion matrix for the multiclass classification task (Twitter). The entries represent the sum on test set examples over 5-fold cross validation.

		Predicted			
		Political	Economic	Religious	Racial
Actual	Political	470	35	11	123
	Economic	16	149	9	9
	Religious	15	4	79	26
	Racial	160	12	37	273

QUALITATIVE ANALYSIS: CASE STUDIES

In Table 5, we present a few examples where our system produced correct/incorrect (top/bottom part) predictions. Figures 5 and 6 present two interesting instances where our multimodal system misclassified.

Table 5: Top: Correctly predicted examples. Bottom: Examples with erroneous predictions.

Post text	OCR Text/Image Description	Actual Class	Predicted Class	Explanation
shabbat shalom to all my jewish friends may the lord bless you	shabbat shalom everyone	Non-Antisemitic	Non-Antisemitic	The terms "friends", "Shabbat", "Shalom" are good clues.
no more jewish wars for israel	I see dead people wherever jews have the power	Antisemitic	Antisemitic	The terms "dead", "jewish" and "wars" are good clues.
Zog (2020): The heartwarming story of a magical dragon who eventually takes control of the entertainment industry.	ZOG (with a picture of a dragon)	Antisemitic	Non-Antisemitic	This post presents a case of sarcasm where ZOG (the dragon cartoon) is used to refer zionist occupied government (ZOG)
Beautiful woman. Not this are zionist woman. They have weapons everywhere.	(No Text)	Racial Antisemitism	Political Antisemitism	The presence of word 'zionist' causes confusion
Banksters jews and the blood from white people	(image with people carrying money bags and dead people)	Economic Antisemitism	Racial Antisemitism	Reference to 'white people' causes confusion.

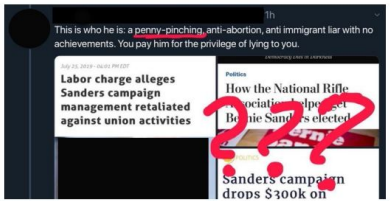


Figure 5: In this example, the image contains screenshot of multiple tweets/posts/articles stitched together posted by someone else.

Figure 6: In this example, the image is a screenshot of a hateful tweet posted by someone else.

LIMITATION & FUTURE WORK

Limitation

Keyword Bias: We observed that posts containing certain keywords like zionists, holocaust, Hitler, Christians, Torah were prone to be classified as antisemitic since majority of training posts containing these keywords were labelled as antisemitic.

Subtlety in the expression of hate: Another set of examples which were misclassified belonged to the category of sarcasm / trolling / subtle hate. It becomes extremely difficult for the system to extract the real intent behind posts expressing views in a subtle manner.

Noise from multiple modalities: Though we showed that adding information from multiple modalities overall helps in antisemitism detection, in a few cases noise present in one of the modalities caused misclassification (as shown in Figure 5 and 6).

Future Work

There are many other forms of hate-speech like Islamophobia, Anti-Asian hate, hate against native community which are yet to be studied in a more robust fashion from a machine learning perspective.

Similar to images, videos have become increasingly common. It will be interesting to develop multimodal systems involving text and videos for detecting antisemitism in the future.

Another interesting direction involves usage of contextual information like user profiles for the classification task.

THANK YOU