

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

Mohit Chandra
International Institute of Information
Technology, Hyderabad
Hyderabad, India
mohit.chandra@research.iiit.ac.in

Dheeraj Pailla*
International Institute of Information
Technology, Hyderabad
Hyderabad, India
dheerajreddy.p@students.iiit.ac.in

Himanshu Bhatia*
International Institute of Information
Technology, Hyderabad
Hyderabad, India
himanshu.bhatia@students.iiit.ac.in

Aadilmehdi Sanchawala
International Institute of Information
Technology, Hyderabad
Hyderabad, India
aadilmehdi.s@students.iiit.ac.in

Manish Gupta†
International Institute of Information
Technology, Hyderabad
Hyderabad, India
manish.gupta@iiit.ac.in

Manish Shrivastava
International Institute of Information
Technology, Hyderabad
Hyderabad, India
m.shrivastava@iiit.ac.in

Ponnurangam Kumaraguru
IIIT, Delhi
New Delhi, India
pk@iiitd.ac.in

ABSTRACT

The exponential rise of online social media has enabled the creation, distribution, and consumption of information at an unprecedented rate. However, it has also led to the burgeoning of various forms of online abuse. Increasing cases of online antisemitism have become one of the major concerns because of its socio-political consequences. Unlike other major forms of online abuse like racism, sexism, etc., online antisemitism has not been studied much from a machine learning perspective. To the best of our knowledge, we present the first work in the direction of automated multimodal detection of online antisemitism. The task poses multiple challenges that include extracting signals across multiple modalities, contextual references, and handling multiple aspects of antisemitism. Unfortunately, there does not exist any publicly available benchmark corpus for this critical task. Hence, we collect and label two datasets with 3,102 and 3,509 social media posts from Twitter and Gab respectively. Further, we present a multimodal deep learning system that detects the presence of antisemitic content and its specific antisemitism category using text and images from posts. We perform an extensive set of experiments on the two datasets to evaluate the efficacy of the proposed system. Finally, we also present a qualitative analysis of our study.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks; Supervised learning by classification**; • **Information systems** → **Social networking sites; Web mining**; • **Social and professional topics** → **Hate speech**.

KEYWORDS

hate speech, antisemitism, multimodal classification, deep learning

1 INTRODUCTION

Online social media (OSM) platforms have gained immense popularity in recent times due to their democratized nature, enabling users to express their views, beliefs, and opinions, and easily share those with millions of people. While these web communities have empowered people to express themselves, there have been growing concerns over the presence of abusive and objectionable content on these platforms. One of the major forms of online abuse which has seen a considerable rise on various online platforms is that of antisemitism.¹

According to International Holocaust Remembrance Alliance (IHRA)², “*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, toward Jewish community institutions and religious facilities.*”. Unlike some forms of hate-speech like sexism, cyberbullying, xenophobia etc., antisemitism originates from multiple aspects. In addition to the discrimination on the basis of race, antisemitism also includes discrimination on the basis of religion (e.g. difference from Christianity), economic activities (e.g. money lending) and political associations (e.g. Israel-Palestine issue, holding power at influential positions).

In recent times, online antisemitism has become one of the most widespread forms of hate-speech on major social media platforms³. This recent trend of increased online hatred against Jews can also be correlated to the increasing real-world crime. According to the Anti-Defamation League’s 2019 report, there has been a 12% jump in the total cases of antisemitism amounting to a total of 2, 107 cases, and a disturbing rise of 56% in antisemitic assaults as compared to

¹<http://www.crif.org/sites/default/fichiers/images/documents/antisemitismreport.pdf>

²<https://www.holocaustremembrance.com/>

³https://ec.europa.eu/information_society/newsroom/image/document/2016-50/factsheet-code-conduct-8_40573.pdf

*Both authors contributed equally to this research.

†The author is also an applied researcher at Microsoft.



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



Text: I see the blews are at it again.

Figure 1: In both these examples, the post text appears to be non-antisemitic, due to absence of an explicit reference to Jews. But when looked along with the image, it can be classified as antisemitic.

2018 across the U.S.⁴ Unlike studies on some major forms of online abuse like racism [1], cyber-bullying [4] and sexism [21], online antisemitism present on the web communities has not been studied in much detail from a machine learning perspective. This calls for studying online antisemitism in greater depth so as to protect the users from online/real world hate crimes.

In previous studies, it has been shown that real-world crimes can be related to incidents in online spaces [39]. Hence, it is essential to build robust systems to reduce the manifestation of heated online debates into real-world hate crimes against Jews. Due to the myriad of content on OSMs, it is nearly impossible to manually segregate the instances of antisemitism, thereby calling for the automation of this moderation process using machine learning techniques. Although antisemitic content detection is such a critical problem, there is no publicly available benchmark labeled dataset for this task. Hence, we gather multimodal data (text and images) from two popular social media platforms, Twitter and Gab.

Oftentimes abusive content flagging policies across various social media platforms are very minimalistic and vague, especially for a specific abuse sub-category, antisemitism. Although various social media platforms have taken measures to curb different forms of hate-speech, more efforts are required to tackle the problem of online antisemitism.⁵ As a result, we provide a detailed categorization methodology and label our datasets conforming to the same. Besides the dataset challenge, building a robust system is arduous because of the multimodal nature of the social media posts. A post with benign text may as well be antisemitic due to a hateful image (as shown in Fig. 1). Thus, it becomes essential to take a more holistic approach rather than just inferring based on text. As a result, we take a multimodal approach which extracts information from text as well as images in this paper.

Fig. 2 shows a detailed architecture of our proposed multimodal antisemitism detection system which leverages the recent progress in deep learning architectures for text and vision. Given a (text, image) pair for a social media post, we use Transformer [35]-based models to encode post text, and Convolutional Neural Networks

(CNNs) to encode the image. Next, we experiment with multiple fusion mechanisms like concatenation, Gated MCB [9], MFAS (Multimodal Fusion Architecture Search) [25] to combine text and image representations. The fused representation is further transformed using a joint encoder. The joint encoded representation is decoded to reconstruct the fused representation and also used to predict presence of antisemitism or an antisemitism category.

We believe that the proposed work can benefit multiple stakeholders which includes – 1) users of various web communities, 2) moderators/owners of social media platforms. Overall, in this paper, we make the following contributions: (1) 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. (2) 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, demonstrating its practical usability. (3) We provide a detailed qualitative study to analyse the limitations and challenges associated with this task and hate speech detection in general.

2 RELATED WORK

In this section, we present the past work, we broadly discuss– (1) studies on antisemitism popular in sociology domain (2) popular hate speech datasets for Twitter and Gab, (3) deep learning studies for broad detection of hateful text content, and (4) overview of applications of multimodal deep learning.

2.1 Previous Studies on Antisemitism

Antisemitism as a social phenomenon has been extensively studied as part of social science literature [28, 29]. These studies have helped to explore the history behind antisemitism in-depth but lack quantitative analyses. Apart from the studies based on Sociology, there have been a few empirical studies. In one of the primary

⁴<https://www.adl.org/news/press-releases/antisemitic-incidents-hit-all-time-high-in-2019>

⁵<https://www.adl.org/holocaust-denial-report-card#the-online-holocaust-denial-report-card-explained->

works, researchers collected around 7 million images and comments from ‘4chan’ and ‘Gab’ to study the escalation and spread of antisemitic memes in a longitudinal study [42]. Ozalp et al. [20] trained statistical machine learning models on data collected from twitter to detect antisemitic content. In contrast to this, our work focuses on detection of antisemitism through a multimodal deep learning framework. Recently, in an another work, researchers focused on detailed annotation analysis based on IHRA’s guidelines for antisemitic content [14]. Our work on the other hand, encompasses the IHRA’s guidelines and incorporates fine-grained classes of antisemitism. Although antisemitism detection is such a critical problem, unfortunately, there hasn’t been any rigorous work on this problem from a deep learning perspective. We fill this gap in this paper.

2.2 Hate Speech Datasets for Twitter and Gab

Hatespeech detection has become a popular area for research and there have been quite a few works on dataset creation. Waseem and Hovy [37] annotated 16,914 tweets, including 3,383 as ‘sexist’, 1,972 as ‘racist’ and 11,559 as ‘neither’. Davidson et al. [5] annotated ~24K tweets for ‘hate speech’, ‘offensive language but not hate’ or ‘neither’. Another recent work presented a dataset comprising of 44,671 posts from various social media platforms and annotated them as offensive or not. Gab was launched in 2016 and hence there are only a few dataset based studies. Qian et al. [26] proposed a hate-speech dataset on Gab with 33,776 posts annotated for hate versus non-hate. Chandra et al. [3] annotated 7,601 Gab posts for ‘Biased Attitude’, ‘Act of Bias and Discrimination’ or ‘Violence and Genocide’. Unlike the previous studies which have focused on general hate-speech or some popular sub-categories (like racism, sexism), we focus specifically on data related to antisemitism.

2.3 Deep Learning Methods for Detection of Various forms of Online Abuse

Deep learning has emerged as one of the most popular methods for hate-speech detection especially in Text-only NLP problems. Founta et al. [8] proposed a Recurrent Neural Networks (RNN) based framework for classification of racism & sexism, offensive speech, and cyberbullying using text and metadata. In contrast to this, we propose a general framework which doesn’t require any metadata. Serrà et al. [31] showed that character level based LSTMs (Long Short-Term Memory networks) can also be effective for abuse classification. In a more recent work, Parikh et al. [22] proposed a neural framework for classifying sexism and misogyny.

Apart from LSTMs, Convolutional Neural Networks (CNNs) have been shown to be fairly successful for this task as they retain the spatial information to extract position invariant features. Gambäck and Sikdar [10] used CNNs to classify the tweets into racist, sexist, both or none. Park and Fung [23] proposed a two-step hybrid approach for classification on hateful text into sexist or racist. They presented a hybrid CNN-based architecture which used sentence and word embeddings. Badjatiya et al. [1] compared multiple deep learning architectures to classify tweets into racist, sexist, or neither. Unfortunately, there has been no previous work on exploring deep learning for antisemitism detection. Also, recently Transformer [35] based methods have shown to outperform traditional deep learning

methods like RNNs and LSTMs. Hence, we resort to methods like Bidirectional Encoder Representations from Transformers (BERT) and Robust BERT Approach (RoBERTa). Besides text, we also leverage semantics extracted from the accompanying image for improved antisemitism detection.

2.4 Applications of Multimodal Deep Learning

With the huge availability of multimodal data, multimodal deep learning has been harnessed to improve the accuracy for various tasks like Visual Question Answering (VQA) [33], fake news/rumour detection [16], etc. Inspired by the success of Optical Character Recognition (OCR) on images for textVQA [33], we also run OCR on the post images and use them for the classification task. Recently, [11, 27, 40] have explored use of multimodal deep learning for general abuse detection from datasets like Reddit + Google images, and Twitter. Our proposed system differs from these previous methods in two important aspects: (1) they use traditional deep learning recurrent text methods like RNNs and LSTMs, while we investigate the application of more promising Transformer-based methods, (2) while previous methods were proposed for general hate, we focus on antisemitism.

3 ANTISEMITISM CATEGORIZATION

Besides annotating every post as antisemitic or not, we also annotate them for finer categories of online antisemitism. While there exists a good amount of literature exploring the ways in which antisemitism manifests itself, we primarily followed the categorization proposed by Brustein [2] as it covers the major aspects of antisemitism. In his work, he explored the history behind the hatred against Jews and has categorized antisemitism into four categories, namely: (1) Political (2) Economic (3) Religious (4) Racial. We augmented this categorization with additional inputs from the detailed IHRA’s⁶ definition. We describe each category of antisemitism in detail in the following.

3.1 Political Antisemitism

Political Antisemitism can be defined as the hatred toward Jews based on the belief that Jews seek national and/or world power. In many of the cases lying in this category, Jews are portrayed to be controlling major political parties, governments, and decision making bodies. We also include the cases where they are accused of controlling media for promoting their interests (printing, Hollywood, etc). Furthermore, in some other cases, Jews are accused of being more loyal to Israel and blamed for the various socio-political crises. For example, *The jews run congress through threats and intimidation*.

3.2 Economic Antisemitism

Economic Antisemitism is based on the implicit belief that Jews perform and control the economic activities which are harmful for others or the society. This notion further exhibits multiple facets like Jews are undeservedly wealthy, greedy, dishonest, materialistic or cheaters. For example, *the driving force behind globalism is jewish finance and greed*.

⁶<https://www.holocaustremembrance.com/working-definition-antisemitism>

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, antisemitism, jewish people, race mixing

3.3 Religious Antisemitism

Religious Antisemitism deals with bias and discrimination against Jews due to their religious belief in Judaism. Cases belonging to this category portray Jews as anti-Christ, Christ-killers, or against the teachings of the *Bible*. Oftentimes, the posts also target Jewish religious institutions as well as their spiritual leader (*Rabbi*). For example, *may god strike down each and every one of these filthy jewish antichrists*.

3.4 Racial Antisemitism

Unlike religious antisemitism, racial antisemitism is based on the prejudice against Jews as a race/ethnic group. Posts belonging to this category display a sense of inferiority for the Jewish race by portraying them as degenerate or attaching certain negative character as naturally inherited by them. Many posts in this category refer to false Jewish conspiracy for racial intermixing and blame them for LGBTQ+ related issues. Along with considering everything which discriminates Jews based on ethnic grounds in this category, we have also included instances talking about *Holocaust* and its denial, since racial prejudice against Jews was one of the main cause for the aforementioned event.⁷ For example, *white is right which makes the jews always wrong*.

4 ONLINE ANTISEMITISM DATASETS

We collect datasets from two popular social media platforms Twitter and Gab. In this section, we discuss details related to data collection, annotation and basic statistics.

4.1 Data Collection

We choose Twitter and Gab as the OSM platforms to gather data for our study. While Twitter has strict anti-abuse policies and an active content moderation team, Gab is an alt-right social media website with relaxed moderation policies. Recently, Gab has gained massive popularity especially among those who have been banned from mainstream web communities for violating their hate speech policy [41]. After gathering a massive collection of posts from Twitter as well as Gab, 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.⁹ This lexicon contains common racial slurs used against Jews along with other words like 'Jewish', 'Hasidic', 'Hebrew', 'Semitic', 'Judaistic', 'israeli', 'yahudi', 'yehudi' to gather non-antisemitic posts as well, thereby helping maintain a balanced class distribution. Presence of these terms does not necessarily indicate presence/ absence of antisemitism and hence we manually annotate the posts.

4.2 Data Annotation

For the annotation task, we selected four undergraduate students who are fluent in English. The annotators were given a detailed guideline along with examples to identify instances of antisemitism. Moreover, to ensure that the annotators had enough understanding of the task, we conducted multiple rounds of test annotations followed by discussions on disagreements. In the annotation procedure, each example was annotated by three annotators and the disagreements were resolved through discussion between all annotators. Each example was 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) if the example is antisemitic then assign the respective antisemitism category. We used Fleiss' Kappa score [7] to compute the inter-annotator agreement. The Fleiss' kappa score came out to be 0.707 which translates to a substantial agreement between the annotators. We removed all user sensitive information and followed other ethical practices to ensure user privacy.

4.3 Data Statistics

Table 1 shows the post distribution across various classes for the two datasets. As observed, 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'.⁸ Table 2 shows the frequent unigrams and bigrams for each of the antisemitism categories. 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,

⁷<https://www.britannica.com/topic/anti-Semitism/Nazi-anti-Semitism-and-the-Holocaust>

⁸https://en.wikipedia.org/wiki/New_antisemitism

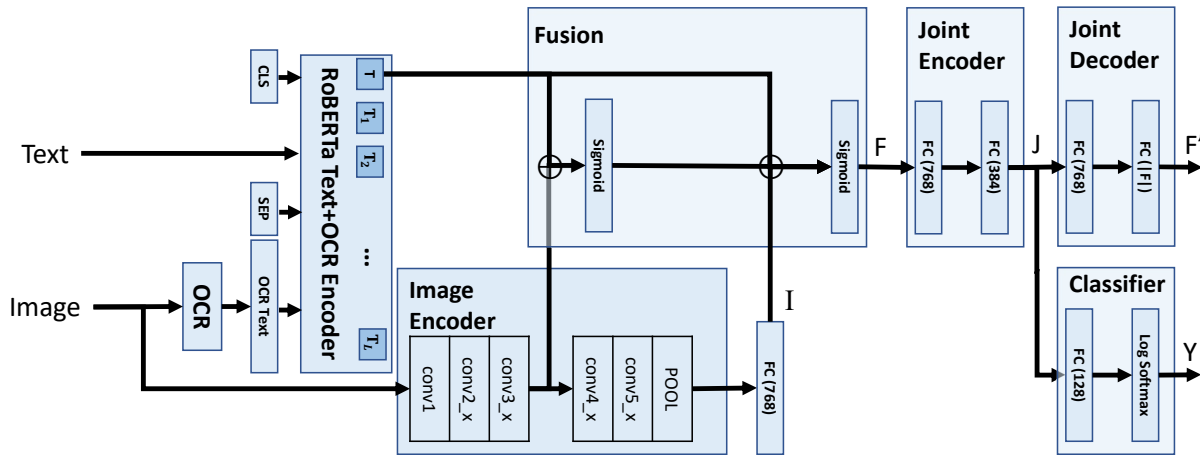


Figure 2: Proposed multimodal system architecture.

after pre-processing for Gab and Twitter respectively. We make the data publicly available.⁹

5 MULTIMODAL ANTISEMITISM CATEGORIZATION SYSTEM

Fig. 2 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. Given a (text, image) pair for a social media post, we use Transformer [35]-based models to encode (1) post text and (2) OCR text extracted from the image. We use CNNs to encode the image. Next, the fusion module combines the text and image representations. The fused representation is further transformed using a joint encoder. The joint encoded representation is decoded to reconstruct the fused representation and also used to predict presence of antisemitism or an antisemitism category. The entire network is trained end-to-end using back-propagation. Since our datasets are relatively small, we fine-tune the pre-trained networks on the presented datasets for multimodal classification. The code for the proposed system can be found here.⁹ We now describe each module in detail.

5.1 Text + OCR Encoder

We remove URLs and non alpha-numeric characters. The cleaned text is tokenized and then encoded using the BERT/RoBERTa tokenizer and encoder respectively [38]. 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). We found the Google’s Vision API to perform the best for the broad range of images we had in the dataset (from newspaper articles to memes). The extracted OCR text goes through the same pre-processing process as the post text. We experiment with BERT and RoBERTa for text encoding since they have been shown to lead to high accuracy across multiple NLP tasks.

BERT [6] is a transformer encoder with 12 layers, 12 attention heads and 768 dimensions. We used the pre-trained model which has been trained on Books Corpus and Wikipedia using the MLM (masked language model) and the next sentence prediction (NSP) loss functions. The input to the BERT model is obtained as a concatenation as follows: (CLS, post text, SEP, OCR text) where CLS and SEP are the standard classification and separator tokens respectively. The 768-dimensional representation T for the “CLS” token from the last encoder layer is used as input by the fusion module.

RoBERTa [19] is a robustly optimized method for pretraining natural language processing (NLP) systems that improves on BERT. RoBERTa was trained with much more data – 160GB of text instead of the 16GB dataset originally used to train BERT. It is also trained for larger number of iterations up to 500K. Further, it uses larger byte-pair encoding (BPE) vocabulary with 50K subword units instead of character-level BPE vocabulary of size 30K used for BERT. Finally, compared to BERT, it removes the next sequence prediction objective from the training procedure, and a dynamically changing masking pattern is applied to the training data.

5.2 Image Encoder

We perform Gaussian normalization for each image, and resize images to 224x224x3 size and feed them to a CNN. We augment our training data using image transformations such as random cropping and random horizontal flipping. We connect the output of second last layer from these networks to a 768 sized dense layer. Output from this layer I is used as the image representation.

Few CNN architectures are popular like: AlexNet [18], InceptionV3 [34], VGGNet-19 [32], Resnet-152 [12] and Densenet-161 [13]. We chose Resnet-152 and Densenet-161 since they have been shown to outperform the other CNN models across multiple vision tasks.

5.3 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 [9] and

⁹<https://github.com/mohit3011/Online-Antisemitism-Detection-Using-MultimodalDeep-Learning>

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

(3) MFAS [25]. Gated MCB (Multimodal Compact Bilinear) pooling combines multimodal features using an outer product followed by a sigmoid non-linearity. We also experimented with Hadamard inner product but found it to be worse compared to gated MCB, in line with previous literature on multimodal deep learning. As shown in Fig. 3 in [25], MFAS (Multimodal Fusion Architecture Search) first concatenates text and image representations from an intermediate hidden layer, applies a sigmoid non-linearity, and then concatenates this with final layer text and image representations along with a sigmoid non-linearity.

5.4 Joint Encoder/Decoder and Classifier

The fused representation F is then passed through a series of Dense layers (768 and 384) to obtain a joint encoded vector J . J is fed to two modules: joint decoder and classifier. The joint decoder again consists of dense layers of sizes 768 and $|F|$. The classifier feeds the output J to a dense layer of size 128 and then finally to the output log-softmax layer. 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.

6 EXPERIMENTS

In this section, we discuss hyper-parameter settings; results using text-only, image-only and multimodal classifiers; and qualitative analysis using attention visualization, error analysis and case studies.

6.1 Hyper-Parameter Settings for Reproducibility

We use the following experimental settings. We perform 5 fold cross validation where we split our labeled data in 64:16:20 as our train, validation, test split for each fold. All hyper-parameters were tuned using validation set. For the MFAS fusion module, we use the block 2 output as the intermediate layer output since it gave us the best results compared to output from other blocks (on validation set). For Gated-MCB all experimental settings were used as suggested by the reference paper. For MFAS, $|F| = |F'| = 2,816$ (which is $3 * 768 + 512$); for other fusion methods $|F| = |F'| = 1,536$ (which is $2 * 768$).

For all experiments, we use Adam optimizer [17]. We experimented with a range of learning rates and found $lr = 2e^{-6}$ as the best one. To improve the stability of the system across the samples

we used a batch normalization layer before the Dense layers. For the Dense layers, we use dropouts with a drop probability of 0.2. We used RELU non-linearity after all our dense layers except the final output layer. We train our system for a max of 50 epochs with early stopping, with a batch size of 4. For all the results, we report 5-fold cross-validation accuracy and macro-F1. For further details of hyper-parameters, we refer the reader to look at our codebase¹⁰.

6.2 Results using Text-only and Image-only Classifiers

We experiment with five popular pre-trained text embedding/network based classifiers and two pre-trained image network classifiers. For the text-only classifiers we use GloVe [24], FastText [15], BERT [6] and RoBERTa [19]. We also experiment with Founta et al. [8]'s method which is an attentional RNN model with GloVe embeddings. For the image only classifiers we experiment with ResNet-152 [12] and DenseNet-161 [13]. We also experimented with VGG-19 but did not see any better results.

Table 3 provides the comparative results. We make the following observations: (1) Compared to the text-only methods, the image-only models provide much lower accuracy. We believe this is because images related to antisemitic posts are usually memes, screenshots, or news articles that usually don't carry any spatial-visual features. (2) Among the text-only classifiers, Transformer based methods performed better than the rather shallow approaches like FastText and GloVe. BERT and RoBERTa lead to very similar results. (3) Among the image-only models, ResNet-152 performs the best except for binary classification on Gab.

6.3 Results using Multimodal Classifiers

In this experiment we tested different fusion mechanism for our proposed multimodal classifier. From Table 3, we observe that ResNet-152 is the best image encoder and RoBERTa is the best text (post text + OCR) encoder. Hence, we perform multimodal experiments with these encoders only. We show the results obtained for this multimodal experiment in Table 4. In addition to this, we also compared the performance of our proposed architecture with the baseline model from [11] (FCM). FCM uses GloVe for text encoder and InceptionV3 for image encoder.

¹⁰<https://github.com/mohit3011/Online-Antisemitism-Detection-Using-MultimodalDeep-Learning>

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

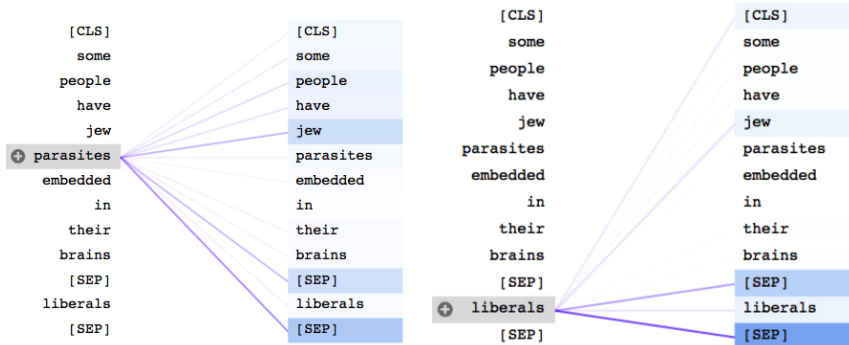


Figure 3: Text + OCR encoder module attention visualization



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

We make the following observations: (1) Each of the three variants of the proposed architecture beat the baseline by a huge margin. (2) Results in Table 4 are much better compared to those in Table 3 except for the binary classification for Twitter. For Gab, we see a massive increase of ~2 and ~4 percentage points in accuracy and F1 for both binary and multi-class tasks respectively. The improvement in results when using both text and images (i.e., across Tables 4 and 3) is better for Gab overall compared to Twitter. This is because the images on Gab are much more rich and informative compared to those on Twitter. (3) MFAS based multimodal fusion approach outperforms the Gated MCB and concatenation based fusion approaches.

6.4 Qualitative analysis: Attention Visualization

To gain better insights into the the proposed system, we visualize attention weights for both the Text + OCR (using bertviz [36]) and the Image encoder (using GradCAM [30]). Figure 3 shows the visualization for a self-attention head from the last encoder layer in the Text + OCR module. 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”. We observe high attention between the word ‘parasites’ with ‘jew’ apart from the standard RoBERTa [SEP] tokens showcasing that the system identified that this text refers Jews as parasites. Similarly, in Fig. 3 (right), the word ‘liberals’ present in the OCR text output shares higher attention

weights with the word ‘jew’ from the post text content showing the cross-attention learnt by the system.

Fig. 4 shows the GradCAM visualization of the image in the post, we observe higher attention on the region of the *Happy Merchant Meme* face which is usually used as a symbol of antisemitism. Another interesting observation is that the text in the image doesn’t get much attention, which makes our choice of adding OCR suitable.

6.5 Qualitative analysis: Error Analysis and Case Studies

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. Each entry in the confusion matrices represents the sum of examples in the test sets over 5-fold cross validation. As observed in Table 6 and Table 8, the classifier has higher percentage of *False Positives* than the *False Negatives* (where the positive class is ‘Antisemitic’). We believe that this was due to many borderline cases which confused the classifier on topics like ‘*Anti-Israel hate*’, ‘*Issue of Israel-Palestine conflict*’ etc. Additionally, from Table 7 and Table 9 we observe that across both the datasets, the classifier is most confused between the ‘Political’ and ‘Racial’ classes. This could be because many politically oriented posts against Jews also used racial prejudices.

Finally, in Table 5, we present a few examples where our system produced correct/incorrect (top/bottom part) predictions. The last

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 Anti-semitism	Political Anti-semitism	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 Anti-semitism	Racial Anti-semitism	Reference to 'white people' causes confusion.

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

Actual		Predicted	
		Non-antisemitic	Antisemitic
	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.

Actual		Predicted			
		Political	Economic	Religious	Racial
	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.

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

"Explanation" column details the plausible reason for the erroneous cases.

Figures 5 and 6 present two interesting instances where our multimodal system misclassifies. The post referred in Figure 5 had the post text as "calling a Jewish man a penny pincher is anti semitic". The post is condemning antisemitic behaviour but the complex structure of text present in the image makes it hard for the model to extract information correctly. Also, the model does not understand that the original tweet was posted by some other user and not this user. Similarly, the post referred in Figure 6 had the post text as

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

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

"@usermention teams up with another antisemite this time a guy who tweeted an image depicting jews as controlling the world and adding that Jewish money crushes the little people". This post reports an antisemitic behaviour by someone else through the screenshot of the tweet. But, due to the absence of any additional context about the image being a screenshot of the tweet from someone else, makes the system to commit the mistake. These two cases help us surface a broader problem with the current systems capturing information from multiple modalities.

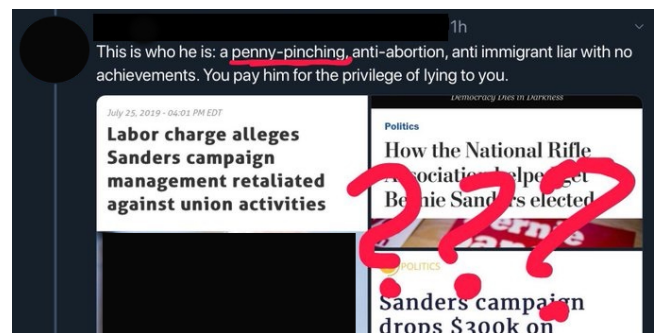


Figure 5: In this example, the image contains screenshot of multiple tweets/posts/articles stitched together posted by someone else.



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

7 LIMITATION

The presented task of antisemitism detection share similar set of limitations as with other hate speech detection tasks like racism/sexism detection. We list a few of the limitations here:

- **Keyword Bias:** As with the other hate speech detection systems, keywords play an important role in the classification of content. 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. This observation is in line with the past research which claims that deep learning models learn this kind of keyword biases.
- **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. This in turn creates a dilemma of freedom of speech/curbing hate speech which is a common problem across various other forms of hate speech.
- **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 Figures 5 and 6).

8 DISCUSSION

In this work we presented the first systematic study on the problem of detection and categorization of antisemitism. We collected and labeled two datasets for antisemitism detection and categorization. We hope that these will accelerate further research in this direction. We proposed a multimodal system which uses text, images and OCR for this task and demonstrated its efficacy on the two datasets. We experimented with single-modal as well as multimodal classifiers and found that combining data from multiple modalities improves the performance and robustness of the system to a small extent for Twitter but massively for Gab. Finally, we also performed a

qualitative analysis of our multimodal system through attention visualisation and error analysis. We observed that the complexities in images and subtlety of hate in text can lead to errors. Images with multiple screenshots, multi-column text and texts expressing irony, sarcasm or indirect references posed problems for the classifier.

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.

REFERENCES

- [1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *WWW*. 759–760.
- [2] William I. Brustein. 2003. *Roots of Hate: Anti-Semitism in Europe before the Holocaust*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511499425>
- [3] Mohit Chandra, Ashwin Pathak, Eesha Dutta, Paryul Jain, Manish Gupta, Manish Shrivastava, and Ponnuram Kumaraguru. 2020. AbuseAnalyzer: Abuse Detection, Severity and Target Prediction for Gab Posts. In *COLING*. 6277–6283.
- [4] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference*. 13–22.
- [5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009* (2017).
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382.
- [8] Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A Unified Deep Learning Architecture for Abuse Detection. In *WebSci*. 105–114.
- [9] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).
- [10] Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 85–90. <https://doi.org/10.18653/v1/W17-3013>
- [11] Raul Gomez, Jaume Gibert, Lluís Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
- [13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jul 2017). <https://doi.org/10.1109/cvpr.2017.243>
- [14] Gunther Jikeli, Damir Cavar, and Daniel Miehling. 2019. Annotating Antisemitic Online Content. Towards an Applicable Definition of Antisemitism. *arXiv preprint arXiv:1910.01214* (2019).
- [15] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).
- [16] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference*. Association for Computing Machinery, 2915–2921.
- [17] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980* [cs.LG]
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.
- [19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692* [cs.CL]
- [20] Sefa Ozalp, Matthew L. Williams, Pete Burnap, Han Liu, and Mohamed Mostafa. 2020. Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society* 6, 2 (2020), 2056305120916850. <https://doi.org/10.1177/2056305120916850> arXiv:<https://doi.org/10.1177/2056305120916850>

- [21] Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label Categorization of Accounts of Sexism using a Neural Framework. In *EMNLP-IJCNLP*. 1642–1652.
- [22] Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing Sexism and Misogyny through Neural Approaches. *Transactions of the Web (TWEB)* (2021).
- [23] Ji Ho Park and Pascale Fung. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. In *Workshop on Abusive Language Online*. 41–45.
- [24] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. 1532–1543.
- [25] Juan-Manuel Perez-Rua, Valentin Vielzeuf, Stephane Pateux, Moez Baccouche, and Frederic Juric. 2019. MFAS: Multimodal Fusion Architecture Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [26] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *EMNLP-IJCNLP*. Association for Computational Linguistics, Hong Kong, China, 4755–4764.
- [27] Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro i Nieto. 2019. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. arXiv:1910.02334 [cs.MM]
- [28] Sarah Salwen. 2010. Antisemitic Myths: A Historical and Contemporary Anthology (review). *Shofar: An Interdisciplinary Journal of Jewish Studies* 27 (01 2010), 176–178.
- [29] Monika Schwarz-Friesel and Jehuda Reinharz. 2017. *Inside the Antisemitic Mind: The Language of Jew-Hatred in Contemporary Germany*. Brandeis University Press.
- [30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.
- [31] Joan Serrà, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali. 2017. Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words. In *Workshop on Abusive Language Online*. 36–40.
- [32] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv* (2014).
- [33] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*. 2818–2826.
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*. 5998–6008.
- [36] Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714* (2019).
- [37] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
- [38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv abs/1910.03771* (2019).
- [39] Dingqi Yang, Terence Heaney, Alberto Tonon, Leye Wang, and Philippe Cudré-Mauroux. 2018. CrimeTelescope: Crime Hotspot Prediction Based on Urban and Social Media Data Fusion. *World Wide Web* 21, 5 (Sept. 2018), 1323–1347.
- [40] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 11–18. <https://doi.org/10.18653/v1/W19-3502>
- [41] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2018. What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *WWW*. 1007–1014.
- [42] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 786–797.