ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Machine Learning for Measuring and Analyzing Online Social Communications

Chris Bronk[1,3],Amaury Lendasse[1,2], Peggy Lindner[1], Dan S. Wallach[3], Barbara Hammer[4]

1- University of Houston - Department of Information and Logistics Technology Houston, USA

2- Arcada University of Applied Sciences - Risklab Helsinki, Finland

3- Rice University - Department of Computer Science Houston , USA

4- Bielefeld University - CITEC - Cognitive Interaction Technology, Germany

Abstract. In this paper, we propose a framework for application of a novel machine learning-based system for analyzing online social communications. As a example, we are targeting anti-Semitic graphical memes posted to social media. We presented very promising preliminary results on a Facebook dataset that consists of a total of 10000 labeled memes. We can conclude that machine learning will soon be able to successfully analyze and monitor complex social communications.

1 Introduction

Internet-delivered social media such as Facebook, Twitter, YouTube, and Instagram allow individuals to create messages in an increasing number of formats and reach incredibly broad audiences with minimal effort and cost. In the last several years, these sites have been employed to influence people with regard to political or social ideologies involving specific and targeted messages [1, 2]. Much of this activity is in the form of memes—graphics that contain images and words designed to attract attention, influence thinking, and motivate action. In the United States and elsewhere, hate groups advocating racist or anti-Semitic action often employ memes to achieve their influence goals.

In studying this problem, we have developed the framework for application of a novel machine learning-based system by which we identify hate speech memes (HSMs), in this case, anti-Semitic graphical memes posted to social media. Such images typically contain well-known images (Star of David, etc.) and text that is anti-Semitic. We then applied our framework to sample meme data provided by social media firm Facebook. What is novel in our approach is the application of machine learning techniques to analyze and classify images and text present in memes to create tools for automated detection.

Our preliminary analysis was aimed at anti-Semitic hate speech memes because they are unfortunately common, typically appear in English, and we can leverage existing image-classification techniques. Neither the graphics alone nor the text alone define a meme being one of hate speech, but the combination of text and image together changes the intrinsic meaning and becomes an artefact

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

of hate speech. Because of this, we face novel challenges in building a suitable classifier. Our techniques may also be applicable to forms of disinformation, including medical advice related to SARS-CoV-2, information regarding elections and election campaigns, and marginalization of other religious groups or minorities.

2 The Internet and Anti-Semitic Hate Speech

"Hate speech" has generated significant interest in contemporary societal discourse within Western democracies as the capacity for individuals to post content to social media has grown. Definitions of this term are present in academic and policy literature as well as law. The Council of Europe, in harmonizing national laws on expression while at the same time recalling the "grave concern about the...resurgence of racism, xenophobia, and anti-Semitism," posited that, "The term 'hate speech' shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance"[3].

A wide variety of research on social media studies Twitter, both because tweets are a short, text format, readily amenable to a variety of natural language processing (NLP) techniques (see, e.g., [4]), and because the vast majority of tweets are public. This contrasts with other social media sites, including Facebook, where most postings are only visible to a user's declared "friends," so would not be easily acquired by running a straightforward web scraping tool. Despite this, a surprising amount of useful information is public on these sites.

Those previous activities as well as those by others have involved the collection of *text* from social media websites. In the race to attract attention and influence individuals online, text is likely to be overshadowed by images and video [5]. For this reason we seek to begin the process of collecting *memes*, i.e., images with overlaid text banners, that may be considered hate speech. We might make this identification through recognition of the text banners, through symbology in the images (e.g., swastikas and related iconography), or through associated metadata (e.g., authorship, posting locations, file names and internal headers, or adjacent online discussions).

Radicalized Anti-Semitic hate speech serves hate groups and marginalizes others [6]. Much of the work in anti-Semitic hate speech discovery undertaken currently is manual in nature [7]. Machine learning tools for that work are immature, however, we can advance the state of the art in the development of tools to automate the process of locating hate speech in the memes of social media.

3 Finding Hate Speech Memes

Social media-enabled hate speech is a moving target. Firms that host Internet content in the United States and elsewhere are sometimes cease hosting controversial speech on their platforms [8]. Closing one venue for hate speech usually

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

spurs the search for a replacement [9]. This constant movement of hateful messaging yields interesting problems for our research efforts.

The creators of hate speech are marketers of ideas in online radical milieus in which, "as specific social environments whose culture, narratives, and symbols shape both individuals and groups, and the social networks and relationships out of which those individuals and groups develop and emerge [10]." For the content creators of hate speech messaging, the challenge is in developing the right images, slogans, and memes to incite action, even violence.

Memes are a photograph or other image typically with some sort of pithy message or slogan appealing to an audience, in our case anti-Semites. The photograph here is of Annelies Marie "Anne" Frank, the Dutch-Jewish diarist who hid from the Netherlands' Nazi occupiers until August 1944, and later died at the Bergen-Belsen concentration camp [11]. The captioned version of the meme is an anti-Semitic "joke" regarding her cremation at Bergen-Belsen. The added text qualifies the modified image as an anti-Semitic HSM [12].

(a) Anne Frank: original image (b) Meme

Fig. 1: Example of how an image can become an anti-Semitic Hate Speech Meme

Our goal is to locate and identify exactly these sorts of anti-Semitic memes. We plan to implement a focused web crawler to collect memes from initially source web pages known to spread hate speech memes. This will prioritize the crawler frontier and manage the hyperlink exploration process [13]. This focused web crawler will be the base crawler which will be combined with a learning-based approach.

4 Identifying HSMs

Locating HSMs is challenging for machines for a variety of reasons. One important issue is how to analyze and extract information to classify images. Secondly, both image features and text both contain important information and they cannot be considered independent. We provide a brief overview of our machine learning approach here, as a more sophisticated description of our machine learning process would be beyond the scope of this brief proposal intended for non-specialist reviewers.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

A common assumption is that the underlying process generating the data is stationary and that the data points are independent and identically distributed (IID). However, contrary to the stationary case, one cannot assume that one can directly employ what has been learned from past data. The model has to keep learning and adapting as new images arrive. Possible ways of doing this include: 1) retraining the model repeatedly on a finite window of past images and 2) using a combination of different models, each of which is specialized on a type of image.

One of the reasons for the problem is the inherent changes in society. As the context is changing, then so is the intrinsic underlying system that produces the images, and the prediction model has to be updated. One way to update the model is to use only a short window of the previous images to build (or train) the model. This task is not trivial since few samples are not sufficient to train an accurate model while too many samples will cover data for a period that cannot be considered stationary.

To classify images, nonlinear classification models are used. We propose to use Deep Belief Networks to extract only necessary information needed to perform an accurate prediction. Though Randomized Neural Networks and Deep Belief Networks are, likely the most suitable machine learning models to classify images, some issues remain to be resolved. First, Randomized Neural Networks are sensitive to the input variables that are used to feed them. We will have to investigate several real-time algorithms to perform the variable selection.

4.1 From Feature Extraction to an Automated Web Crawler

Feature extraction is a critical element in many machine learning and data science applications. It is common to first extract low-level or high-level semantic features from images, text, or other input, and subsequently feed these features into a statistical model [14, 15]. The quality of the extracted memes' features will be a greater determinant of our success than any subsequent analysis or modeling decision towards an automated web crawler and will have to be done in carefully orchestrated ways.

4.1.1 Extraction of local image features using Harris-Laplace Detector

If the classification task to be performed involves images, one type of features that has to be used should extract information about the images themselves. For non image-related classification tasks, other features (like the text features introduced in the next Section) should be used instead.
Local image features are meaningful or informative regions of an image [16]. Think about about a picture of an airplane in the sky — a patch of a uniform blue sky is not very informative, whereas a patch containing the airplane is. Local features usually contain corners, edges or strong changes in color and contrast [17]. These features are specifically made to be invariant to image transformations (scaling, rotation) and noise (for instance from different image encodings). Thus they are useful for finding *similar objects* in different im-

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

ages [18]. These features can also be used for image classification, by finding image patches similar to those from some particular classes [19]. Edge and corner detection in an image uses the derivative (or gradient) of image intensity map, the pixel values of a gray scale image. This is performed using the Harris corner detector [17]. A good overview and other feature detection methods is given in the following summary papers [17].

One commonly used method is SIFT (Scale-Invariant Feature Transform), a local image feature descriptor, based on histograms of oriented gradients (HOG) [20]. Based on our previous research [21], SIFT are sufficient for our needs; relevant libraries already exist and don't have to be implemented from scratch.

4.1.2 Text Features

Owing to the small amount of text embedded in the memes means working with very little text content and some of this text may require understanding of idiom or context. We need to identify text features (TeFe) which can be combined with the image features (IF). As a first step, we will deploy large-scale text extraction on the collected memes using a text detection and recognition approach [22]. The extracted text fragments will be preprocessed and segmented into words (tokenized). Once the tokens are produced, we feed them into a pipeline to create different short text feature representations in the form of term vector matrices which will be utilized during the the machine learning. To build the term vector matrices, we plan to deploy at least two methods: one using a bag-of-words (BoW) approach and the other one using a sparse vector construction, i.e. term frequency-inverse document frequency ($TF \times IDF$). BoW simply counts how many times a word appears in a document and produces wordcount vectors which can be organized as term frequencies (TF). However, the BoW model might suffer from a rarity issue as the rareness of a term is not considered and we are expecting that the text fragments from the memes will contain some rare terms. A weighted approach such as ($TF \times IDF$) will mitigate the issue.

The $TF \times IDF$ scheme originally developed in [23] gives a weight to denote the importance of a term from two aspects: frequency of the term (TF) in a document and inverse document frequency (IDF) in the corpus. The $TF \times IDF$ weight denotes a degree of a term's relevance to a document while discriminating the document from other documents in the corpus. So, the term weight for a document is computed by

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{1}$$

where $tf_{i,j}$ denotes the number of occurrences of term i in j, df_i the number of documents containing i and N the total number of documents.

4.1.3 Metadata Features

Metadata on websites provide information about a page's structure or content and is used primarily by search engines and social networks to better understand

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

webpages. We can utilize metadata primarily to feed the automated web crawler to detect image based content [24] but also as an additional feature for the identification of hate speech memes. The extraction of basic website metadata can be achieved through utilization of a web scraping toolkits such as Scrapy [25] which produces structured feature data sets.

Social media platform metadata contains much more detailed information about relationships between people. They are after all, designed for this purpose. Consequently, social media metadata is even more revealing than basic website metadata in terms of showing the strengths of connections between individuals and groups. We can use those as features to aid in tracing the sources and spread of hate speech memes.

4.2 Building a growing database and adaptive model

When we began work on this idea, no large data set of labeled HSMs existed. Evidence of the importance of this research topic emerged in May 2020 when Facebook launched the "Hateful Memes Challenge and data set for research on harmful multimodal content" [26]. Facebook created a dataset specifically to help AI researchers develop new systems to identify multimodal hate speech. This content combines different modalities, such as text and images, making it difficult for machines to understand. We will employ this dataset to initialize our methodology. Some of the useful categories for us in the challenge are: Race, Ethnicity, Religion and Nationality.

5 Learning from the Facebook Data

The Facebook dataset consists of a total of 10000 labeled memes. The dataset is splitted into a training (8500 memes) and validation set (1500 memes). Our exploratory analysis was done on the training set (C1 – Hate Memes 36%/ C0 – Clean Memes 64%) using only the text features. We first extracted 9379 words. We removed 4434 words that occurred only once since manual inspection showed that these words didn't seem seem to be correlated with C0 or C1.

We choose to a staged approach. We first built a *basic classifier* based on word frequency.

Algorithm 1 Stage 1

1: If Meme x includes Word "y", what is the probability to belong to C1.
2: For each word, we search the Memes that include that word and how many of these Memes belong to C1
3: Find threshold

Using a simple threshold we identified 40 words that the classifier will use.

In *stage 2* we focused on the 200 most frequent words and built a criteria-based classifier. Using an Extreme Learning Machine (ELM, [27]) we built 100k classifiers by selecting 30 words randomly and ranked them based on the following criteria:

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

a For C0, at least 64% of corrected classification (to be as good as NC)

b For C1, the percentage of correct classification pC1 is maximized

In the 3rd stage we find the best classifier using the following algorithm:

Algorithm 2 Stage 3

1: For C0, the percentage of correct classification is computed as $pC0$
2: For C1, the percentage of correct classification is computed as $pC1$
3: $(pC0 + pC1)/2$ is maximized

We achieved the following results: pC0 = 0.86%, pC1 = 0.37% and (pC0 + pC1)/2 = 0.61%

6 Conclusion and Further Work

Through our exploratory work with the Facebook dataset we showed that a relative simple algorithm can lead the way to classification of memes just based on text. It will be important to build machine learning tools that combine analysis of both images and text to determine the semantic content of images. While there are many potential applications of such tools, HSMs are a blunt form of social media tools designed to misinform individuals and dehumanize segments of the population. Making automated tools that can detect such speech may have utility in assuring civility in discourse online without infringing on speech liberties.

References

[1] Alexandra A Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Measuring the prevalence of online hate speech, with an application to the 2016 us election, 2018.

[2] L. Hellmueller, V. Hase, and P. Lindner. Terrorist organizations in the news: A computational approach to measure media attention toward terrorism. *Mass Communication and Society*, 0(0):1–24, 2021.

[3] Council of Europe. *Recommendation, N. O. R (97) 20 Of the Committee of Ministers to member states on "hate speech"*.

[4] Rob Procter, Helena Webb, Marina Jirotka, Pete Burnap, William Housley, Adam Edwards, and Matt Williams. A Study of Cyber Hate on Twitter with Implications for Social Media Governance Strategies. *arXiv e-prints*, page arXiv:1908.11732, Aug 2019.

[5] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2:335, 2012.

[6] R. Willison and M. Warkentin. The expanded security action cycle: A temporal analysis 'left of bang'. In A. Vance, editor, *Proceedings of the Dewald Roode Information Security Workshop*, pages 392–438. International Federation for Information Processing, 2010.

[7] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics, 2012.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

[8] Evelyn Aswad. The role of us technology companies as enforcers of europe's new internet hate speech ban. *HRLR Online*, 1:1, 2016.

[9] Matthew Costello, James Hawdon, and Thomas N Ratliff. Confronting online extremism: The effect of self-help, collective efficacy, and guardianship on being a target for hate speech. *Social Science Computer Review*, 35(5):587–605, 2017.

[10] Maura Conway. From al-zarqawi to al-awlaki: The emergence and development of an online radical milieu. *CTX: Combating Terrorism Exchange*, 2(4):12–22, 2012.

[11] Anne Frank. *Diary of a young girl*. Enrich Spot Limited, 2016.

[12] Stevie Voogt. Countering far-right recruitment online: Cape's practitioner experience. *Journal of policing, intelligence and counter terrorism*, 12(1):34–46, 2017.

[13] C. Saini and V. Arora. Information retrieval in web crawling: A survey. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2635–2643, 2016.

[14] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn van Dolen. Multimodal Popularity Prediction of Brand-related Social Media Posts. In *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, pages 197–201, Amsterdam, The Netherlands, 2016. ACM Press.

[15] Quinten McNamara, Alejandro de la Vega, and Tal Yarkoni. Developing a comprehensive framework for multimodal feature extraction. *CoRR*, abs/1702.06151, 2017.

[16] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proc. of the 9th European conference on Computer Vision*, pages 404–417, 2006.

[17] Tinne Tuytelaars and Krystian Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.

[18] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[19] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of Nearest-Neighbor based image classification. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, number I, pages 1–8. IEEE, June 2008.

[20] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proc. of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:886–893, 2005.

[21] A. Akusok, Y. Miche, J. Karhunen, K.-M. Bjork, Rui Nian, and A. Lendasse. Arbitrary category classification of websites based on image content. *IEEE Computational Intelligence Magazine*, 10(2):30–41, May 2015.

[22] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD'18, pages 71–79, New York, NY, USA, 2018. Association for Computing Machinery.

[23] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988.

[24] Turek W., Opalinski A., and Kisiel-Dorohinicki M. Extensible web crawler – towards multimedia material analysis. volume 149. Springer, Berlin, Heidelberg, 2011.

[25] Dimitrios Kouzis-Loukas. *Learning Scrapy*. Packt Publishing Ltd, 2016.

[26] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

[27] Yoan Miche, Antti Sorjamaa, Patrick Bas, Olli Simula, Christian Jutten, and Amaury Lendasse. Op-elm: Optimally pruned extreme learning machine. *IEEE Transactions on Neural Networks*, 21(1):158–162, 2010.