A Concept for Automated Polarized Web Content Annotation based on Multimodal Active Learning

Marek Herde[1[0000−0003−4908−122X]], Denis Huseljic[1[0000−0001−6207−1494]], Jelena Mitrović[2[0000−0003−3220−8749]], Michael Granitzer[2[0000−0003−3566−5507]], and Bernhard Sick[1[0000−0001−9467−656X]]

[1] University of Kassel, Wilhelmshöher Allee 73, 34121 Kassel, Germany
{marek.herde | dhuseljic | bsick}@uni-kassel.de
[2] University of Passau, Innstrasse 43, 94032 Passau, Germany
{michael.granitzer | jelena.mitrovic}@uni-passau.de

Abstract. Active learning (AL) techniques hardly cope with complex annotations tasks, where, for example, annotations might express relationships across data modalities. As a use case, we consider the task of automatically detecting and reporting multimodal, polarized web content (PWC). Samples of this content type emerge dynamically, covering a broad spectrum of topics. Thus, training machine learning systems for detecting PWC is challenging, particularly if it needs to be done with minimum annotation cost. In this article, we propose the concept of multimodal AL for complex annotations in the context of PWC detection and formulate the resulting challenges as questions for future research.

Keywords: Active Learning · Multimodal Data · Semantic Annotation · Polarized Web Content · Hateful Memes.

1 Motivation

Supervised *machine learning* (ML) relies on vast amounts of annotated data often provided by human annotators in a labor-intensive process. *Active learning* (AL) addresses this problem of costly data annotation by intelligently querying annotators [2]. The goal is to maximize an ML system's performance while minimizing the annotation cost. Although AL techniques have shown their benefit for classification and regression tasks [7], they hardly cope with more complex annotation tasks, where annotations might

- express relationships across data modalities (A1),
- describe (semantic) relationships between concepts (A2),
- come along with a high level of error-proneness and potential disagreement among annotators due to an ambiguous context (A3),
- or require modeling background knowledge and sociodemographic factors of annotators to estimate the quality of annotations (A4).

As a use case, we consider the task of automatically detecting and reporting potential multimodal [9], abusive web content in political communication, which is in most cases strongly polarized. We use *polarized web content* (PWC) instead of related expressions such as hateful memes [4,5] to highlight this polarized nature. Generally, PWC comes in many forms, is subjective, depends on the context, and frequently requires background knowledge to be understood [13]. In this article, we refer to PWC as multimodal online content, mainly text and images, that can be found on social media and has, e.g., defamatory or abusive characteristics (at least from the viewpoint of certain groups of persons). The left side of Fig. 1 shows a PWC sample composed of an image of the burning World Trade Center on 09/11 and an image of a Muslim congresswoman, Mrs. Ilhan Abdullahi Omar. These two images are combined with a textual contradiction of "never forget" and "you have forgotten". The polarized context arises from combining images and text (A1), which relates the concepts Twin Towers to Muslims and terrorism (A2). Identifying this polarization requires knowledge about American history and politics (A4) or otherwise may result in erroneous annotations (A3). Such PWC samples emerge dynamically and unforeseeably, covering a broad spectrum of concepts. Thus, training ML systems for detecting PWC is challenging, particularly if it needs to be done annotation cost-efficiently.

Within this article, we view PWC detection as a challenging sample application with real-world impact [11] to initiate research on extending AL systems toward complex annotations of multimodal data. Therefore, we propose our concept of *multimodal active learning for complex annotations* (MALCOM) and formulate the associated challenges as questions for future research.

2 Concept

We envision MALCOM as an extension of traditional AL [2], which assumes a single omniscient annotator providing categorical labels as annotations, toward (1) *semantic annotation graphs* (SAGs) [15] as complex, multimodal annotations and (2) an AL strategy selecting pairs of annotators and queries, e.g., samples. The objective is to semi-automatically build models that can identify PWC and analyze it by annotating a potential PWC sample with an SAG. Such an SAG describes the PWC samples' contents, explains why its contents can be seen as polarized, and reflects the potential uncertainty in that analysis. Fig. 1 shows a PWC sample and its SAG to illustrate this objective. In the following, we outline our two envisioned extensions of AL and PWC detection in more detail.

Extension 1 – Complex, Multimodal Annotations: Existing PWC detection approaches focus on standard supervised learning settings with categorical labels as annotations [1,6,16]. The outputs or embeddings of vision and language models are typically combined as input for a final decision model. Our proposed SAGs represent an alternative combination strategy for the two modalities of images and text. SAGs allow decisions on a higher semantic level, which fosters explainability and decouples objective annotation tasks such as concept analysis of images and texts from more subjective decisions on polarization. We

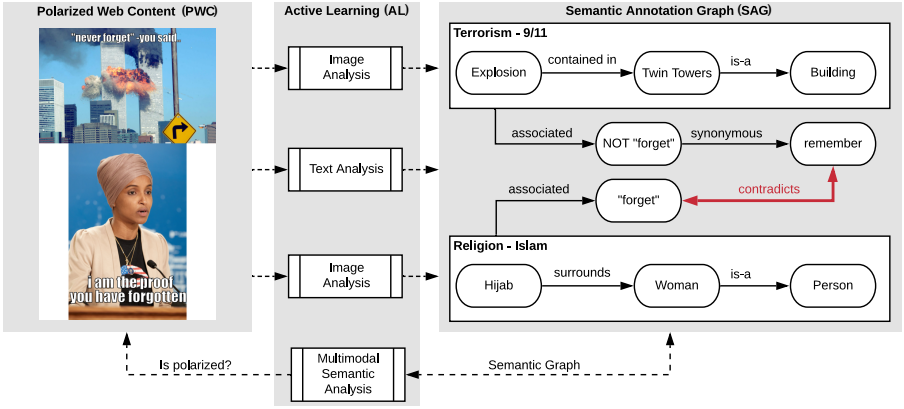

Fig. 1. PWC sample[1] with racist motive (left) and corresponding SAG (right) obtained by combined image and text analysis: Rounded rectangles represent concepts, arrows represent relations, and rectangular boxes represent inferred concepts. As a typical indicator of PWC, a contradicting relation is highlighted in red. AL (center) is applied for (1) unimodal image and text analysis and for (2) inferring whether a sample is polarized from the SAG through multimodal semantic analysis. In this simplified figure, we do not show additional information that is provided with the SAG, e.g., uncertainty regarding object classes or positions in images, relations beyond contradictions, etc.

argue that this is a more efficient way of generating precise automatic classifications of PWC. Methodologically, we have to go far beyond annotating images or text individually but considering their relationships. Annotations may describe positions of objects in images (regions of interest), comparisons of two images or texts, the importance of specific contexts for decisions, a degree of polarization, confidence estimates regarding decisions, etc. We need to develop a proper semantic model, e.g., ontologies [8,12], covering the different modalities and being understandable for annotators. This also includes the ability to express very different PWC concepts over different modalities that go beyond contradictions but include more fuzzy concepts such as antitheses or correlations between concepts.

Extension 2 – Query and Annotator Selection: Identifying PWC requires contextual knowledge of (very recent) events, e.g., pandemics [14]. So instead of building one generic model, we aim at building specialized models for different kinds of PWC, which use pre-trained models (per modality), and fine-tune them in an AL cycle. Extending the AL cycle towards complex annotations of multimodal data, as sketched in Fig. 2, starts with the question of integrating different modalities. First, we consider a pool of annotated unimodal data, i.e., texts and images, which we use to create unimodal models that can annotate

[1] Image above is a compilation of assets, including ©Getty Images/Spencer Platt and ©Getty Images/Adam Bettcher, used under the "Hateful Memes Dataset License Agreement". It is taken from "The Hateful Memes Challenge" [5] for illustrative purposes only and any person depicted in the content is a model.

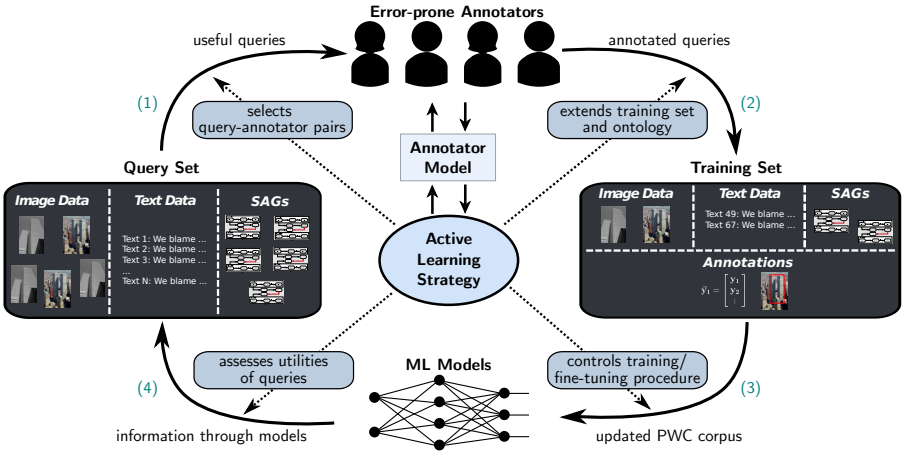

Fig. 2. AL cycle for MALCOM with four main steps: (1) Useful queries are selected from a set of all possible queries regarding potential PWC. For example, we may query annotations for the objects in an image or ask whether an SAG is polarized. (2) Selected queries are presented to a subset of annotators with possibly different (e.g., educational) backgrounds. This subset is determined through an ML-based annotator model estimating the annotators' qualifications. Subsequently, the annotated queries update the training set. (3) The training set representing the current PWC corpus is used to (re-)train several ML models, e.g., an object detection model. (4) The trained models provide information regarding the query set such that the AL cycle starts again using this information for query selection.

the unimodal data semantically. This process in each case results in an SAG, i.e., a typed, attributed graph defined through an ontology-based annotation scheme. Later, the SAGs are merged into a joint, multimodal SAG. Similar to traditional AL strategies, we need to identify promising candidates – initially images and texts, later multimodal SAGs – to be annotated. To consider the problem's multimodal nature, the annotations' semantic properties, and the annotators' diverse backgrounds, we must develop new AL selection strategies that account not only for the respective data sample but also for the different kinds of queries and the qualifications of certain annotators regarding the PWC sample at hand. These qualifications (also referred to as annotator performance [2]) may depend on various aspects such as the respective PWC category (e.g., politics) or educational background (e.g., Master's degree in political sciences). The annotator model predicting such qualifications needs to be sensitive to annotator minorities, e.g., by estimating similarities between annotators. Otherwise, we risk ignoring annotator minorities' opinions regarding PWC. Moreover, we must consider that answers regarding the degree to which content is polarized may be highly subjective, i.e., uncertain from an ML perspective [3]. Establishing an objective definition of PWC, similar to hate speech research [10], is a possible way of reducing the subjectivity of PWC annotation.

3 Research Questions

We conclude this article with the following six research questions derived from the above key research objective and the required extensions.

- How can we define ontology-based annotation schemes to express a human's reasoning over classifying web content as (gradually) polarized or not?
- How can we extract image descriptions (part of the SAG) from potentially polarized images (part of the PWC) considering different uncertainty types?
- How can we extend AL for object detection in potentially polarized images?
- How can we extend AL over text extracted from the images to identify rhetorical figures and automatically analyze textual content to create semantic annotations automatically?
- How can we merge unimodal SAGs and extend AL to train models, e.g., graph convolutional networks [17], assessing PWC via multimodal SAGs?
- How can we evaluate the above techniques and build or extend data corpora [5] for research?

Acknowledgements

SPONSORED BY THE
Federal Ministry of Education and Research

The project on which this article is based was partly funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049. The authors are responsible for the content of this publication. Furthermore, the authors thank Chandana Priya Nivarthi, Stephan Vogt, Mohammad Wazed Ali, and the anonymous reviewers for their insightful comments to improve this article.

References

1. Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring Hate Speech Detection in Multimodal Publications. In: WACV. pp. 1470–1478. Snowmass Village, CO (2020)
2. Herde, M., Huseljic, D., Sick, B., Calma, A.: A Survey on Cost Types, Interaction Schemes, and Annotator Performance Models in Selection Algorithms for Active Learning in Classification. IEEE Access **9**, 166970–166989 (2021)
3. Huseljic, D., Sick, B., Herde, M., Kottke, D.: Separation of Aleatoric and Epistemic Uncertainty in Deterministic Deep Neural Networks. In: ICPR. pp. 9172–9179. Virtual (2021)
4. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C.A., Bull, P., Lipstein, G., Nelli, T., Zhu, R., et al.: The Hateful Memes Challenge: Competition Report. In: NeurIPS 2020 Competition and Demonstration Track. pp. 344–360. Virtual (2021)
5. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In: NeurIPS. pp. 2611–2624. Virtual (2020)

6. Kumar, A., Sachdeva, N.: Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. Multimed. Syst. (2021)
7. Kumar, P., Gupta, A.: Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey. JCST **35**(4), 913–945 (2020)
8. Kühn, R., Mitrović, J., Granitzer, M.: GRhOOT: Ontology of Rhetorical Figures in German. In: LREC. Marseille, France (2022)
9. Lahat, D., Adali, T., Jutten, C.: Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. Proceedings of the IEEE **103**(9), 1449–1477 (2015)
10. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PLOS ONE **14**(8), 1–16 (2019)
11. Mishra, P., Yannakoudakis, H., Shutova, E.: Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. arXiv:1908.06024 (2019)
12. Mitrović, J., O'Reilly, C., Mladenović, M., Handschuh, S.: Ontological representations of rhetorical figures for argument mining. Argument & Computat. **8**(3), 267–287 (2017)
13. Sood, S.O., Antin, J., Churchill, E.: Using Crowdsourcing to Improve Profanity Detection. In: AAAI Spring Symposium 2012 – Wisdom of the Crowd. pp. 69–74. Palo Alto, CA (2012)
14. Uyheng, J., Carley, K.M.: Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. JCSS **3**(2), 445–468 (2020)
15. Vidal, J.C., Lama, M., Otero-García, E., Bugarín, A.: Graph-based semantic annotation for enriching educational content with linked data. KBS **55**, 29–42 (2014)
16. Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., Predovic, G.: Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification. In: ALW. pp. 11–18. Florence, Italy (2019)
17. Zhang, S., Tong, H., Xu, J., Maciejewski, R.: Graph convolutional networks: a comprehensive review. Comput. Soc. Netw. **6**(1), 1–23 (2019)