Data & Policy (2024), 6: e54
doi:10.1017/dap.2024.44

CAMBRIDGE
UNIVERSITY PRESS

COMMENTARY

Understanding to intervene: The codesign of text classifiers with peace practitioners

Julie Hawke[1], Helena Puig Larrauri[2], Andrew Sutjahjo[3] and Benjamin Cerigo[3]

[1]University of Notre Dame; Build Up, South Bend, USA
[2]Build Up, London, UK
[3]datavaluepeople, Amsterdam, Netherlands
Corresponding author: Julie Hawke; Email: jhawke@nd.edu

Received: 01 October 2023; **Revised:** 30 July 2024; **Accepted:** 23 August 2024

Keywords: data science; large language models; machine learning; peacebuilding; social media

Abstract

Originating from a unique partnership between data scientists (datavaluepeople) and peacebuilders (Build Up), this commentary explores an innovative methodology to overcome key challenges in social media analysis by developing customized text classifiers through a participatory design approach, engaging both peace practitioners and data scientists. It advocates for researchers to focus on developing frameworks that prioritize being usable and participatory in field settings, rather than perfect in simulation. Focusing on a case study investigating the polarization within online Christian communities in the United States, we outline a testing process with a dataset consisting of 8954 tweets and 10,034 Facebook posts to experiment with active learning methodologies aimed at enhancing the efficiency and accuracy of text classification. This commentary demonstrates that the inclusion of domain expertise from peace practitioners significantly refines the design and performance of text classifiers, enabling a deeper comprehension of digital conflicts. This collaborative framework seeks to transition from a data-rich, analysis-poor scenario to one where data-driven insights robustly inform peacebuilding interventions.

Policy Significance Statement

This commentary offers a practical experiment to better harness the power of data in addressing digital conflicts and advancing peacebuilding efforts. By uniting data scientists and peace practitioners in a participatory approach, we empower policymakers with actionable insights. The emphasis on usability and real-world applicability ensures that our methodologies can inform policy decisions effectively. In an era marked by digital polarization, understanding online communities is crucial. Our case study on online Christian communities in the United States serves as a model for similar analyses across diverse contexts. Policymakers can use our findings to craft interventions that mitigate conflict and promote dialogue. This collaborative framework represents a shift from data abundance to informed action, aligning with the needs of policymakers striving for evidence-based solutions in the digital age.

1. The Problem

Social media has transformed the way people communicate with each other, leading to shifts across wide areas of research and policy discussion that continue to bring bedrock democratic principles of speech,

https://doi.org/10.1017/dap.2024.44 Published online by Cambridge University Press

privacy, and protection to bear. This transformation has especially impacted the field of peacebuilding. Peacebuilding is a practice that aims to resolve or transform conflicts by addressing both immediate and systemic forms of violence. It involves a wide range of activities, including conflict prevention, negotiation, social justice advocacy, and the promotion of social cohesion. Peacebuilders work at various levels, from the interpersonal to the international, and employ a wide range of strategies to meet these goals. Increasingly, peacebuilders also feel compelled to engage online, where digital media is exacerbating societal divides in communities across the world (Kubin and Von Sikorski 2021; Lorenz-Spreen et al. 2021; Schirch 2021; Stray et al. 2023). While it is not the only or even primary conflict driver, patterns of digital content consumption and interaction are intertwined with polarization, antidemocratic practices, dehumanization, and violence (Bilewicz and Soral 2020; Silverman et al. 2020; King et al. 2017; Dangerous Speech Project 2023; Bail et al. 2018). The problem of online polarization is widespread and indiscriminate, impacting media users directly with broken relationships, harassment, and manipulation as well as indirectly with eroded institutions, tension, and violence regardless of individual media use. On the other hand, social media is concurrently instrumental in both formal and everyday forms of peacebuilding. It also enables coordination, connection, empathy, and even increased political knowledge and participation (Alexander 2014; Houston et al. 2015; Lorenz-Spreen et al. 2021). In this complex milieu, peace practitioners, mediators, and policymakers have the remit to reduce violence and advance social justice and cohesion.

Picture a scenario in a community where tensions between different identity groups have been escalating. The conflict escalates both online and offline reciprocally. Insults, hate speech, and even calls to violence spread through social media, making people feel less safe or trusting at school or work. Fights or confrontations are filmed and shared online, diffusing the conflict further. Misinformation and sensationalism flood the information environment. Aware of these dynamics, peacebuilders aim to design an intervention. To find the most strategic entry point and to ensure their intervention does not make things worse, they conduct a conflict analysis that includes information on key stakeholders, socio-political and economic factors, triggering events, key narratives about the "other side," and their current patterns of interaction. This could lead to communication campaigns, workshops with community leaders, dialog or deliberation groups, mediation between group representatives, or interventions to reduce online hostility.

All of these programs, and many others not listed, have been implemented as a response to increased polarization signaled by online interaction. Peacebuilders and mediators conducting conflict analyses are increasingly interested in looking at social media platforms as sites in which conflicts are articulated and enacted. However, while the relevance and impact of online discourse are known, the systematic and data-heavy nature of the problem make it challenging to establish clear links between online behaviors and offline conflict dynamics. For instance, does increased hostile rhetoric online precede physical violence, or vice versa? When do certain types of online discourse correlate with specific offline conflict behaviors? Answering such questions is becoming more important for those engaging in conflict prevention and response. However, wrangling big data is prohibitive for community organizations and activists and elusive even to supporting conflict researchers that opt for qualitative methods to capture the complexity and granularity of online interactions.

In this commentary, we first outline the problem faced by peacebuilders in navigating the complex, data-rich but analysis-poor environment of conflict on social media. We then present our position advocating for the participatory design of custom text classifiers—automated tools that categorize text into predefined groups based on their content—that leverage both domain expertise and supervised learning strategies to enable more effective and actionable conflict analysis. Next, we delve into the technical limitations of existing approaches and introduce a test aimed at overcoming these challenges—an active learning methodology in which we train a simple classifier to determine polarization as an annotator is annotating the dataset (Cohn et al. 1996). Finally, we discuss broader implications for research and practice.

When conducting a social media analysis for peacebuilding, there is a persistent technical challenge: of the text snippets scraped or consumed from application programming interface (APIs), only a small percentage are relevant for the research priority, in this case, the specific conflict dynamics or divisive behaviors under investigation. For example, if a peacebuilder is analyzing online discourse related to a particular ethnic or religious conflict, they may be interested in identifying content that expresses hate speech, dehumanization, or other forms of polarizing language between the groups involved. However, the vast majority of the collected data may not directly pertain to these specific aspects of the conflict. The challenge lies not just in reducing a large dataset (potentially millions of text snippets) to a manageable subset of relevant content (perhaps tens of thousands), but in doing so in a way that captures the nuances and complexities specific to the conflict at hand. While efficient data processing is a crucial component in this process of narrowing, techniques to do so are readily available. Ensuring the quality, contextual relevance, and completeness of the analysis, however, presents a more complex challenge. If we claim to have found 50k relevant results, it is difficult to ascertain whether this figure is accurate or if there were actually 100k relevant snippets in our dataset, but we only managed to identify 50k of them. This highlights the interplay between technical tradeoffs and analytical challenges. We must balance precision (the proportion of identified snippets that are truly relevant) with recall (the proportion of snippets that we successfully identified as relevant from all the relevant snippets), all while maintaining a context-sensitive analysis that truly reflects the complexities of the conflict under study.

It is possible to use various pre-existing natural language processing techniques like generalized text classification, sentiment, or named entity recognition from the outset.[1] However, there are also limitations. These range from simple techniques, which are transparent and easy to interpret, to more complex approaches. Bag-of-words, for example, represent text as a collection of word frequencies, making it straightforward to understand which words contribute to classifications. Other methods include TF-IDF weighting, which considers word importance, and *n*-gram models that capture short phrases. More advanced techniques like word embeddings and neural networks can capture subtle semantic relationships but are less transparent. Transformer models, such as those used in GPT, are now the most widely known by nondata scientists, yet are fully opaque—black boxes, due to the mass of data represented in their neural network. Generalized text classification, such as hate speech detection, can give a sense of conflict-related content in a large collection of texts (corpus) but can also miss key features of the relevant text snippets that are specific to a conflict. The context also influences what hate speech or negative sentiment is. For example, a generalized hate speech detector, which is trained on a broad range of data, is good at finding general hate speech but is not able to find context-specific or coded hate speech. For example, it would naturally miss locally used euphemisms or seemingly neutral terms that have acquired hateful connotations within a specific context. Finally, it is generally not possible to configure these generalized text classifications in understandable ways that can produce results that are intelligible to peacebuilders while also maintaining an acceptable level of performance. The complex inner workings of these models, such as the specific linguistic features or patterns they rely on to make classifications, are opaque or difficult to interpret, even for technical experts. Additionally, the outputs of these models may not align with the specific categories or dimensions of analysis that are most relevant to peacebuilders' work, which can limit their practical utility.

Peace practitioners therefore find themselves amid a data-rich but analysis-poor environment. This makes the analysis of digital conflicts poor or incomplete, which impacts the design of conflict interventions or policy recommendations.

2. The Position

To design evidence-based programs or policy responses, peacebuilders need a fine-grained understanding of the divides playing out on digital media, how they are spreading, who is involved, and their real-world

[1] Text classification: A process in machine learning where a model is trained to categorize pieces of text into predefined groups based on their content. For example, sorting comments into categories such as "hate speech" or "non-hate speech."

Sentiment Analysis: A computational technique used to determine the emotional tone behind a body of text. This is used to understand the attitudes, opinions, and emotions expressed in written language, classifying them as positive, negative, or neutral.

Named Entity Recognition (NER): A process in natural language processing that identifies and classifies key information in text into predefined categories. Examples include identifying names of people, places, organizations, dates, and other specific data.

https://doi.org/10.1017/dap.2024.44 Published online by Cambridge University Press

impacts. On its face, this is a disputed position. Hirblinger et al. (2024) argue for a critical engagement with the "assumption that uncertainty in peace processes can simply be overcome through better data and evidence supplied with the help of digital technologies—and indeed, if seeking certainty is always necessary or even desirable." However, the premise underlying our position is not a positivist uncovering of causal linkages. We argue that better data, especially social media data, illuminates the complexity and multiplicity of contexts in ways that support critical engagement. Crucially, this requires involvement with the data-gathering process itself. At the same time, to produce actionable analysis, researchers or data scientists need to use custom text classifiers that better reflect significance to a context within datasets that are too large for manual review. Partnership with domain experts in the participatory design of text classification is needed. Furthermore, it is still unknown how OpenAI's closed-source large language models (LLMs) and other recently released open-source LLMs can perform for the specific text classifications that a peacebuilder might need, nor what methodologies yield the best results when using LLMs to meet the needs of peacebuilders.

The remainder of this article outlines how we sought to solve this problem by combining practitioner knowledge with different forms of AI to improve the ease and quality of text classifications. The application is specific to peacebuilding practice, but it extends to other areas of policy and practice that require engagement with the inherently emotional, subversive, and/or nuanced content that characterizes conflict discourse.

The authors are from a uniquely positioned partnership across data scientists (datavaluepeople) and peacebuilders (Build Up) who are building open-source social media analysis tools that can be used across diverse contexts to inform programming. A critical issue with AI deployment is the absence of tailor-made applications codesigned with practitioners through a "human-in-the-loop" approach, leveraging their contextual expertise and grounded objectives. Beyond demonstrating this experiment, we advocate for researchers to focus on developing frameworks or models that prioritize being usable and participatory in field settings, rather than perfect in simulation under ideal, predefined conditions for testing.

3. Understanding the Challenge: Technical Limitations

The work outlined in this article builds on previous work conducted by datavaluepeople and Build Up to develop text classifiers that identify topics in social media content relevant to a conflict setting using a combination of lists of keywords that map to classes, annotated examples, and custom-built machine learning models.

A simple, effective, and understandable method for peacebuilders to find relevant data is to use a list of keywords that are mapped to classes. Simply, if a text snippet contains a keyword, it is then classified with the mapped class. Although this seems effective, it misses complexities within the language and can give many false positives and false negatives. This method was used in a social media listening program deployed to support eighteen civil society organizations across six Arabic-speaking countries (British Council 2022). To address these limitations and improve the accuracy of text classification, machine learning models have been explored as an alternative approach.

Methodologies to classify relevant data using machine learning models can outperform keyword-list methods but face challenges of an imbalanced dataset. In previous projects, of the 2000–5000 examples that we can reliably get our annotators to label, we expect the prevalence of relevance class to be 20–100 text snippets. As a conventional rule rooted in practical experience, creating a reliable classifier based on examples requires a minimum of 100 examples to ensure sufficient data to capture the variability within the class and minimize overfitting, particularly in cases of imbalanced datasets. Therefore, if the prevalence of relevant snippets is consistently at the lower end of our observed range, a domain expert may need to annotate up to 10,000 snippets to gather sufficient examples for effective model training. Because annotators—typically domain experts or peace practitioners—have other professional commitments and limited time to dedicate to the annotation task, the question of efficiency is in support of, not counter to, goals of quality and completeness in analysis. Because capturing nuance and complexity is dependent on their context-sensitive knowledge, it is important to maximize the value of their input while minimizing the time required.

The new methodology described in this article would reduce the number of examples that need to be annotated while also reducing the miss rate, thus decreasing the time needed to get to a working classifier. Previous research has shown that generative AI could successfully support classification models (Gilardi et al. 2023; Møller et al. 2023; Rathje et al., 2024; Törnberg 2023). Still, it is unknown to what extent and with what level of training or prompt engineering this could include conflict-related data.

4. Addressing the Challenge: Experimenting with Text Classifiers for Conflict Analysis in the United States

For this test case, we focused on how Christian identity as expressed on social media is furthering polarization of attitudes towards others and mistrust of others in the United States. This case is a useful reference for this commentary because it demonstrates the nuanced and sensitive nature of conflict-related data. Faith communities are highly important organizing spaces in the United States and are generally a key component of both bonding and bridging social capital, providing a sense of unity and shared values that strengthen bonds within communities and foster collaboration across diverse groups (Putnam 2000). With the rise of Christian nationalism and the recent violence it has engendered, these same ideological or cultural mechanisms are risk factors for radical attitudes, intentions, and behaviors. The majority of faith leaders may not call themselves peacebuilders, but peacebuilders view faith leaders who prioritize peacemaking and social justice roles as integral actors for conflict transformation (Sampson 2007). Currently, these concerned faith leaders are either in the dark about the types of content and ideological narratives their constituents are espousing online, or they are seeing the issues and tension destructive social media interactions are causing but do not know if, when, and how to translate their positions of trust or authority into online communities. Leaders who are trying to "plugin" and challenge risky dynamics online are using positive messaging tactics that are too diffuse and abstract to address the real problems, that is a "love your neighbor" meme, and can even contribute to further inflaming the divisions. Alternatively, others have had their positions put in jeopardy for addressing more specific tensions around racism, belonging, or governance.

Our theory of change is that if we could support local faith leaders to understand the social media map and patterns of interaction within their online faith communities and facilitate the subsequent design and implementation of responsive pilot actions (both on and offline) that address polarizing dynamics, leaders would be equipped to understand and respond to online tensions and their offline impacts. However, the time-intensive nature of annotation, the process of manually labeling or categorizing data to train machine learning models, poses a significant barrier to participation for interlocutors with demanding schedules and varying levels of interest in the inner workings of data science, despite the known benefits of participation in improving data analysis and subsequent actions. To address this barrier, our experiment focuses on reducing the annotation time required to build a robust classification model. By minimizing the time commitment needed for annotation, we aim to make participatory methods more accessible and feasible for a wider range of stakeholders. As such, this experiment was collaborative but not participatory without the inclusion of a wider range of stakeholders. Two authors, as peacebuilding experts, developed the classification guidelines and annotations, with one author providing additional validation based on their contemporaneous role as local faith leader. The other two, as data scientists, developed the experiment methodology.

Our previous participatory research on conflict drivers in social media has led to the development of a framework for categorizing divisive online behavior (Puig Larrauri 2023). This framework distinguishes between deliberate tactics used to harass or manipulate people on social media and contextual changes resulting from the amplification of such tactics. In the current test case, contextual changes, specifically attitude polarization, are the primary driver of an emerging conflict. Attitude polarization is defined as perceptual shifts toward stereotyping, vilification, dehumanization, or deindividuation of others (Hawke et al. 2022). To analyze this conflict driver, we propose using two text classifiers to narrow down relevant

https://doi.org/10.1017/dap.2024.44 Published online by Cambridge University Press

social media content: one to identify content pertaining to the group(s) involved in the conflict dynamic (in this case, Christian identity) and another to identify content expressing or contributing to attitude polarization. Although specific to this test case, we argue that this dual-classifier approach, focusing on group identity and conflict-relevant behavior, is essential for peacebuilders to effectively navigate social media landscapes and draw meaningful conclusions for conflict analysis.

The annotation guidelines for these two text classifiers are as follows:

Relevant: references Christianity

To classify a piece of text as referencing Christianity, consider the following guidelines. Please keep in mind the goal is to minimize false positives and avoid misclassification with other religions:

1. Christian Groups, Denominations, or Identities: Does the text mention specific Christian groups or denominations such as Catholics, Protestants, Evangelicals, or Orthodox Christians? Note that generic terms like “believers,” “followers,” or “church” may apply to other religions as well and should not be used as sole indicators of Christianity.
2. Scripture or Quotes from Christian Leaders: Does the text reference or quote the Bible, including specific books or verses? Does it quote recognized Christian leaders like Pope Francis, Martin Luther, Billy Graham, etc.? Be careful with general religious or spiritual quotes that could be from any religious tradition.
3. Christian Theology, Rituals, or Practices: Does the text mention specific rituals and practices unique to Christianity, such as baptism, holy communion, or confirmation? Or theological concepts unique to Christianity like the Trinity, Original Sin, Salvation, etc.? Be aware that rituals such as prayer, fasting, or giving alms are common to many religions and should not be considered definitive indicators of Christianity.
4. Christian Symbols and Figures: Are there references to Jesus Christ, Virgin Mary, the Cross, the Resurrection, or other figures and symbols unique to Christianity? Remember that some figures (like Jesus and Mary) are also recognized in other religions, so context is essential.
5. Christian Holidays or Observances: Does the text mention Christian holidays such as Christmas, Easter, Good Friday, Lent, or Pentecost? These are unique to Christianity and can be considered strong indicators.
6. Context: Overall context is extremely important in this classification. Words like “church” could refer to a Christian place of worship or to a group or organization in a nonreligious sense. Similarly, references to Jesus or Mary could be from a Christian context, or they could be from a Muslim context, as both figures are recognized in Islam. Always consider the overall context of the text when making decisions.
7. Doubtful Cases: If the classification of a text as referencing Christianity is not clear, it is preferable to err on the side of caution and not classify it as such. It is also recommended to set up a “maybe” category for texts where the context is uncertain.

Remember, only texts that clearly reference Christianity should be classified as such. If a text could equally apply to another religion, it should not be classified as referencing Christianity. Always consider the context and the overall meaning of the text, and not just individual words or phrases.

Attitude polarization: perceptual shifts towards stereotype, vilification, dehumanization, or deindividuation of others	To classify a piece of text as reflecting attitude polarization, consider the following guidelines: 1. Stereotyping: Does the text generalize a specific group of individuals, attributing certain characteristics to all members of the group regardless of individual differences? Stereotyping often reduces complex individuals to simple, monolithic representations. 2. Vilification: Does the text defame or demonize a particular group, person, or entity, inciting fear? This could be through exaggeration, misrepresentation, or biased framing that presents the subject in a negative, harmful light. 3. Dehumanization: Is the text stripping a group or individual of their human qualities or personality? Dehumanization can be seen in language that compares people to animals, machines, or objects, or that otherwise denies their humanity, dignity, or individuality. 4. Deindividuation: Does the text reduce individuals to anonymous members of a group, ignoring their unique characteristics or personal identities? Deindividuation often involves erasing individuality to emphasize group identity, sometimes with the implication that all members of the group are interchangeable or identical. 5. Extreme Language and Absolutism: Does the text use extreme language or make absolute statements? Attitude polarization often involves language that is absolutist (e.g., "always," "never"), extreme (e.g., "worst," "best"), or dichotomous (e.g., "us versus them," "right versus wrong"). 6. Lack of Empathy or Understanding: Does the text lack empathy or understanding for other perspectives or experiences? This could involve ignoring or dismissing the viewpoints or experiences of others, or showing a lack of willingness to understand or empathize with them. 7. Invalidation: Does the text invalidate people's identity's existence? 8. Doubtful Cases: If the classification of a text as reflecting attitude polarization is not clear, it is preferable to err on the side of caution and not classify it as such. It is also recommended to set up a "maybe" category for texts where the context is uncertain. Remember, only texts that clearly reflect attitude polarization should be classified as such. Always consider the context and the overall meaning of the text, not just individual words or phrases.

5. Experiments Methodology

The data for this test is drawn from Twitter and Facebook, utilizing their respective APIs. On Twitter, the focus was on 20 influencers who frequently manifest Christian identity in their discourse, with a bifurcation based on two spectrums: on the x-axis of polarizing or pluralizing, and a y-axis of conservative or liberal. For Facebook, the approach entailed identifying 10 phrases commonly employed by individuals to express Christian identity. Utilizing these phrases, a search was conducted to identify topperforming public pages that employ these phrases, with the stipulation that the page administrators are based in the United States. The final dataset iterated from these starting points was comprised of 8954 tweets and 10,034 Facebook posts.

Our experiments are focused on reducing the amount of time an annotator needs to spend to get enough annotations to make a classification model (approximately 10,000 annotations are needed to garner enough positive instances to create a model from a random sample). We employ a methodology called active learning, in which we train a simple classifier to determine polarization as an annotator is annotating the dataset (Cohn et al. 1996). Instead of training a model on a randomly selected subset of data, active learning is a type of human-in-the-loop approach that focuses on areas where the model is uncertain or where additional information could significantly improve the model's performance. Dor et al. (2020) describe the need for active learning within a "challenging coupled setup, frequently encountered by real-world users—where labeled data is scarce, and the prior of the desired class is small." In essence, text classifiers need a large amount of labeled data for model training, but labeling data is especially difficult when the specific category of interest is not common in the data. Simply labeling a random sample, in this case, would not guarantee that there would be enough positive examples to train the classifier effectively. Meeting the priority of efficiency, this approach iteratively refines a model by incorporating the most useful data points for a machine learning model to be able to capture the distinguishing properties of defined classes into the training process. Instead of having humans label a large, random sample of data, active learning directs human effort towards the most ambiguous or uncertain instances. The following steps are used:

First, we train a simple bag-of-words (BOW) model with a linear classifier to make initial predictions.[2] This classifier classifies all the text snippets in the dataset, and its output determines the selection of specific snippets for human annotation to make the classifier better.[3] Furthermore, after each sample that the human annotates, the model's representation of the specified class, in this case what makes something polarizing, is changed to include the new information that it gives. The simple classifier is updated, or retrained if you will, after each annotated text snippet. This represents an incremental learning strategy, rather than the more traditionally used batched learning methods (Schlimmer and Granger 1986). To determine which text samples to present for annotation, we employ Gaussian sampling based on the model's certainty scores. The probability of a sample being selected is inversely proportional to the model's confidence in its classification. The model assigns a score between 0.0 and 1.0 to each sample, where 0.0 indicates absolute certainty that the sample does not belong to the target class (e.g., not affective polarization), and 1.0 signifies complete certainty that it does (e.g., is affective polarization). This sampling method prioritizes uncertain examples for annotation while occasionally including high-confidence samples to check for errors, balancing the exploration of uncertain areas with the verification of high-confidence predictions. Because of the retraining process, the probability of each sample being seen also gets updated based on new information of each annotation, which speeds up the search for the most advantageous examples to give the human annotator.

Upon completion of this phase, we discard the bag-of-words model itself but retain the annotations as the primary input into the development of a LLM.[4] This is done in a single run without employing further active learning. The LLM offers the advantages of being more efficient in identifying affective polarization instances because of its improved context awareness (containing representations of words, subwords, and textual context) compared to keyword-based annotation methods. We then verify the model performance using the f1 score[5] as well as by providing a sample set to the practitioner for a manual

[2] We used the spacy.TextCatBOW.v3 architecture for the original classifier (Explosion, 2024 *textcat.py*, GitHub).

[3] While we did not use this strategy for this experiment, using a list of keywords to seed the initial annotation space is argued to provide the benefit of getting to a good model quicker.

[4] We use BERT (Bidirectional Encoder Representations from Transformers), a transformer-based machine learning technique for natural language processing (Devlin et al. 2018).

[5] The f1 score is a metric that balances precision (accuracy of positive predictions) and recall (completeness of positive predictions). Ranging from 0 to 1, it provides a single measure of a classification model's performance. It is a suitable metric in instances where datasets are imbalanced.

https://doi.org/10.1017/dap.2024.44 Published online by Cambridge University Press

check of where a model's predictions differ from the annotator's labels, including cases of both false positives, false negatives, and uncertain classifications. If the initial model proves adequate, we are able to apply it to the full dataset. If not, we proceed again to annotate additional text that the model has difficulty with or is uncertain about.

This approach can encounter a significant challenge when transitioning from a BOW model to an LLM. Unlike BOW models, which can be rapidly updated, LLMs require substantial time for retraining and a nontrivial period for classifying new texts. The iterative process of updating model weights, reclassifying samples, and selecting new annotation candidates becomes prohibitively time-consuming with LLMs when using the active learning approach.

To address computational constraints associated with using LLMs in combination with active learning, we precompute the LLM's internal representations of text snippets in our dataset. These precomputed representations are then fed into a simpler machine learning classifier, which serves as the active learning vehicle. This approach allows us to maintain the efficiency of single-text snippet cycles in the active learning loop, as the simpler classifier can be rapidly updated and queried. The model in the loop can thus run on the annotator's local computer, facilitating immediate updates and selections after each annotation. While it forgoes the ability for LLM fine-tuning, this hybrid solution enables a cost-effective and responsive methodology for the participatory design of text classifiers, directly addressing the key challenges outlined in the introduction to the problem. The following step is to retrain the LLM using the full set of annotations acquired during both the initial phase and the active learning process using precomputed LLM representations. As before, we discard the interim model used in the active learning loop, focusing solely on the annotations it helped generate. This retraining step allows us to leverage the context-aware advantages of the LLM while incorporating the context-specific insights gained through the annotation process. As before, we verify the updated model by manually checking a sample set based on an assessment of its performance using the f1 score. If the model is sufficient, we apply it to the full dataset. If not, we repeat the process, iterating through the annotation and refinement cycle as needed.

The iterative methodology outlined enables the development of cost-effective, context-specific text classifiers through participatory design, addressing the key challenges outlined in the introduction and balancing efficiency, accuracy, and domain expertise in peacebuilding applications.

6. Implications for Research and Practice

The approach and motivation described in this commentary demonstrate the potential for the collaborative development of text classifiers that integrate domain expertise with advanced AI techniques. As the field of peacebuilding increasingly recognizes the impact of social media on conflict dynamics, researchers and practitioners must work together to develop tools and methods that can effectively analyze and interpret online discourse. Similar collaborative efforts have been undertaken in related domains. For example, an approach combining experts from nongovernmental organizations (NGOs) with AI models has been used to build hate speech detectors and response support (Tekiroglu et al. 2020; Chung et al. 2021). These kinds of experiments, which focus on complex datasets, minimal time investment from domain experts, and the intelligibility of the classifier-building process, provide a framework for future research in this area.

While the results of our test are preliminary and ongoing, the process highlights the importance of involving practitioners in the design and validation of these tools to ensure their relevance and usefulness in real-world contexts. Over the course of this year, we will also run two additional experiments—one with data from Nigeria, and another with data from across the Sahel—re-using the same experiment methodology while defining context-specific classifiers and annotation guidelines. These experiments are partly research-validation work but are also being directly implemented into practice as they are embedded in ongoing peacebuilding programs run by Build Up.

Our focus is to bridge research directly into practice by conducting rigorous and replicable experiments into customizable text classifiers that also respect three critical success criteria for peacebuilding practice. First, we experiment with complex datasets, meaning they contain the nuanced, context-specific language, and dynamics that characterize real-world conflicts. These datasets are directly related to actual

https://doi.org/10.1017/dap.2024.44 Published online by Cambridge University Press

conflicts of interest to peacebuilding practitioners, as they reflect the challenges and subtleties that peacebuilders encounter in their work. Second, the process of building a text classifier requires a minimal amount of time and effort from peacebuilders. Finally, the process for building a text classifier is intelligible to peacebuilders, such that they can understand the effect of choices and definitions on the resulting classifier, and therefore its impact on the conflict analysis they can subsequently conduct. Collaboration with practitioners is not viewed as a means to an end toward developing autonomous models that could claim to interpret conflict contexts independently, obscuring bias and projecting false infallibility. Rather, this approach serves to externalize a peacebuilder's interpretation of a conflict context onto a machine learning model, acknowledging the inherently subjective nature of those judgments and making them explicit. In theory, unlimited time and data could more fully mitigate the known and unknown biases of both people and models through diverse sampling and model refinement. Practical constraints, however, necessitate balancing model performance against resource investment—a tradeoff requiring honest assessments of practitioner objectives and text classification outcomes.

This collaboration between domain experts, technical experts, and platforms underscores the potential for enhanced model development and fine-tuning when grounded in practical, on-the-field insights. This is what we need more of: collaboration that is central at the experimental phase will lead to better deployment outcomes—in peacebuilding and comparable fields where practical application is essential.

Data availability statement. None declared.

Author contribution. Conceptualization: Puig Larrauri; Hawke. Methodology: Cerigo; Sutjahjo. Data curation: Cerigo. Writing original draft: Hawke; Puig Larrauri. Writing review and editing: Cerigo; Sutjahjo. All authors approved the final submitted draft.

Funding statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None declared.

References

Alexander DE (2014) Social media in disaster risk reduction and crisis management. *Science and Engineering Ethics 20*(3), 717–733. https://doi.org/10.1007/s11948-013-9502-z

Bail CA, **Argyle LP**, **Brown TW**, **Bumpus JP**, **Chen H**, **Hunzaker MFB**, **Lee J**, **Mann M**, **Merhout F and Volfovsky A** (2018) Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences 115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bilewicz M and Soral W (2020) Hate speech epidemic: the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology 41*, 3–33. https://doi.org/10.1111/pops.12652

British Council (2022) Digital Maps Reports. Available at https://howtobuildup.org/wp-content/uploads/2023/02/DMaps_Report_2022.pdf (Retrieved 29 September 2023).

Chung YL, **Tekiroglu SS and Guerini M** (2021) Towards knowledge-grounded counter narrative generation for hate speech. arXiv preprint arXiv:2106.11783.

Cohn DA, **Ghahramani Z and Jordan MI**. (1996) Active learning with statistical models. *Journal of Artificial Intelligence Research 4*, 129–145.

Dangerous Speech Project. (2023) Dangerous Speech: A Practical Guide. Available at https://dangerousspeech.org/guide/ (retrieved 14 April 2024).

Devlin J, **Chang M-W**, **Lee K and Toutanova K** (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dor LE, **Halfon A**, **Gera A**, **Shnarch E**, **Dankin L**, **Choshen L**, **Danilevsky M**, **Aharonov R**, **Katz Y and Slonim N** (2020) *Active Learning for BERT: An Empirical Study.* In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp. 7949–7962.

Explosion. textcat.py. GitHub repository. Available at https://github.com/explosion/spaCy/blob/master/spacy/ml/models/textcat.py (accessed 30 July 2024).

Gilardi F, **Alizadeh M and Kubli M** (2023) Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.

Hawke J (2022) Archetypes of Polarization on Social Media. Build Up Blog. Available at https://howtobuildup.medium.com/archetypes-of-polarization-on-social-media-d56d4374fb25 (retrieved 29 September 2023)

Hirblinger AT, **Wählisch M**, **Keator K**, **McNaboe C**, **Duursma A**, **Karlsrud J**, **Sticher V**, **Verjee A**, **Kyselova T**, **Kwaja CMA**, **Perera S** (May 2024) Forum: making peace with un-certainty: reflections on the role of digital technology in peace processes beyond the data hype, *International Studies Perspectives 25*(2), 185–225. https://doi.org/10.1093/isp/ekad004

Houston JB, **Hawthorne J**, **Perreault MF**, **Park EH**, **Goldstein Hode M**, **Halliwell MR**, **Turner McGowen SE**, **Davis R**, **Vaid S**, **McElderry JA and Griffith SA** (2015) Social media and disasters: A functional framework for social media use in disaster planning, response, and research. *Disasters 39*(1), 1–22. https://doi.org/10.1111/disa.12092

King G, **Pan J and Roberts ME** (2017) How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review 111*(3), 484–501. https://doi.org/10.1017/S0003055417000144

Kubin E and Von Sikorski C (2021) The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association 45*(3), 188–206. https://doi.org/10.1080/23808985.2021.1974395

Lorenz-Spreen P, **Oswald L**, **Lewandowsky S and Hertwig R** (2021) Digital media and democracy: a systematic review of causal and correlational evidence worldwide. Preprint at SocArXiv. https://doi.org/10.31235/osf.io/p3z9v

Møller AG, **Dalsgaard JA**, **Pera A and Aiello LM** (2023) Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. arXiv preprint arXiv:2304.13861.

Puig Larrauri H (2023) How to Find Evidence of Divisive Behavior on Social Media. Build Up Blog. Available at https://howtobuildup.medium.com/how-to-find-evidence-of-divisive-behavior-on-social-media-7b5322d9d65b (retrieved 29 September 2023)

Rathje S, **Mirea D-M**, **Sucholutsky I**, **Marjieh R**, **Robertson C and Van Bavel JJ** (2024) GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences 121* (34) https://doi.org/10.31234/osf.io/sekf5

Sampson C (2007) Religion and peacebuilding. In Zartman IW and Rasmussen JL (eds), *Peacemaking in International Conflict: Methods and Techniques*. United States Institute of Peace Press, pp. 273–323.

Schirch L. (ed.) (2021) *Social Media Impacts on Conflict and Democracy: The Techtonic Shift*. London: Routledge.

Schlimmer JC, **Granger RH** Incremental learning from noisy data. *Machine Learnning 1*, 317–354 (1986). https://doi.org/10.1007/BF00116895

Silverman C, **Lytvynenko J and Kung W** (2020, January 7) Disinformation for Hire: How a New Breed of PR Firms Is Selling Lies Online. BuzzFeed News. Available at https://www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms (retrieved 17 April 2023).

Stray J, **Iyer R and Puig Larrauri H** (2023) The algorithmic management of polarization and violence on social media. Preprint at SocArXiv.

Tekiroglu SS, **Chung YL, and Guerini M** (2020) Generating counter narratives against online hate speech: Data and strategies. arXiv preprint arXiv:2004.04216.

Törnberg P. (2023) Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588.

Cite this article: Hawke J, Puig Larrauri H, Sutjahjo A and Cerigo B (2024). Understanding to intervene: The codesign of text classifiers with peace practitioners. *Data & Policy*, 6: e54. doi:10.1017/dap.2024.44