DIVERSE: A Dataset of YouTube Video Comment Stances with a Data Programming Model

Iain J. Cruickshank
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
icruicks@andrew.cmu.edu

Lynnette Hui Xian Ng
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
lynnetteng@cmu.edu

Amir Soofi
University of California, Los Angeles
Los Angeles, California, USA
soofi@cs.ucla.edu

***Abstract*—Public opinion of military organizations plays a crucial role in their ability to recruit talented individuals. As recruitment increasingly extends into digital spaces like social media, assessing the stance of social media users toward online military content becomes essential. However, there is a notable lack of data for analyzing opinions on military recruitment efforts online, compounded by challenges in stance labeling, which is key to understanding public perceptions. Despite the importance of stance analysis for effective online military recruitment, generating human-annotated, in-domain stance labels is resource-intensive. In this paper, we address both the challenges of stance labeling and the scarcity of data on public opinions of online military recruitment by introducing and releasing the DIVERSE dataset[1]. This dataset comprises comments from videos on the U.S. Army's official YouTube channel. We employed a state-of-the-art weak supervision approach, leveraging large language models to label the stance of each comment toward its respective video and the U.S. Army. Our findings indicate that the U.S. Army's videos began attracting a substantial number of comments post-2021, with a generally balanced stance distribution among supportive, oppositional, and neutral comments, though with a slight skew towards oppositional over supportive comments.**

***Index Terms*—stance detection, social media, large language models, ensemble learning**

I. Introduction

The U.S. military is currently facing a recruitment crisis. Over the past several years, organizations like the U.S. Army have struggled to meet recruitment targets necessary to maintain their desired force structure (Brown, 2023). Social media plays a critical role in understanding public stances (AlDayel and Magdy, 2021) and in influencing recruitment (Ng and Cruickshank, 2023). Therefore, it is essential to understand how military messaging is received on social media, as these perceptions can shape recruiting strategies. More broadly, the ability to assess how social media content resonates with a target audience is crucial for any marketing effort, whether it involves military recruitment, organizational outreach, or consumer influence.

One key indicator of social media content reception is *stance*. Stance detection, which involves identifying an expressed or implied opinion toward a target, remains central to many computational social studies. Various methods and datasets have been developed for stance detection (Allaway and McKeown, 2023; Du Bois, 2007). However, existing datasets typically focus on stance as expressed on a single social media platform, primarily Twitter (formerly X), and usually assign a single stance per comment (Sobhani et al., 2017; Conforti et al., 2020; Mohammad et al., 2016; Hossain et al., 2020; Kochkina et al., 2018; Villa-Cox et al., 2020). The task of stance detection is challenging due to the context-dependent nature of stance (e.g., the need for a target to define stance) and varying definitions of stance itself (Ng and Carley, 2022). As a result, these stance datasets rely on human annotation for label creation, which is prone to error and limits dataset size to the thousands, given the resource-intensive nature of human labeling.

To address these limitations in existing stance datasets, we introduce a new stance dataset based on YouTube comments, labeled through machine-assisted weak supervision. This dataset comprises comments posted on videos from the U.S. Army's official YouTube channel[2]. Comments are initially annotated with weak labels representing stance-related attributes, such as hate speech, sarcasm, and sentiment. These labels are refined by prompting three different Large Language Models (LLMs), and the weak labels are then combined to determine the final stance labels of comments towards the U.S. Army and its content. Through this dataset, we aim to gain deeper insights into public opinion on the U.S. Army and its recruitment efforts.

Our contributions in this paper are as follows:

- **A novel benchmark stance dataset of YouTube comments.** We present a dataset for stance detection that differs from prior benchmarks in several ways. While most stance datasets draw on Twitter posts, ours focuses on YouTube comments, broadening the range of platforms used in stance research. Additionally, unlike most datasets, ours provides stance labels for multiple targets within a comment (similar to Sobhani et al. (2017)) and includes weak labels for contentious issues, conspiracy theories, and other emotional content. Finally, our dataset exceeds the size of the largest current benchmark dataset by over three times (Conforti et al. (2020) with approximately $32,000$ comments, versus more than $216,000$ in our dataset).

[1]https://doi.org/10.5281/zenodo.10493803

[2]Official U.S. Army YouTube Channel: https://www.youtube.com/@usarmy

- **A data-programming-based stance-labeling methodology.** We leverage recent advances in creating large, high-quality labeled datasets with contextual knowledge to generate reliable stance labels Ratner et al. (2019); Huang et al. (2024); Smith et al. (2024). Our approach uses weak labels, including hate speech detection, sentiment analysis, sarcasm identification, keyword presence, and LLM evaluations, combined through weak supervision to produce the final stance labels.

II. Related Work

Stance Datasets: The development of stance detection has led to several datasets. Ng and Carley (2022) consolidated datasets with social media posts spanning multiple topics, primarily labeled as favor, against, or neither, representing sentiment toward a target (Mohammad et al., 2016). Some datasets incorporate implicit and explicit support or denial to capture the degree of stance expression (Qazvinian et al., 2011). These datasets have been used to train algorithms, including support vector machines (Elfardy and Diab, 2016), logistic regression models (Augenstein et al., 2016), neural networks (Fang et al., 2019; Siddiqua et al., 2019), and deep learning models (Küçük and Can, 2020). Recently, Large Language Models (LLMs) have also been utilized for stance detection (Zhang et al., 2023a; Gatto et al., 2023).

YouTube Datasets: While YouTube data has been studied for influence campaigns (Marcoux et al., 2021; Hussain et al., 2018), conspiracy theories (Liaw et al., 2023), and toxicity (Obadimu et al., 2019), few datasets have been publicly released, especially for stance detection on YouTube comments. The existing research often excludes dataset publication due to proprietary constraints (Liaw et al., 2023). Although commercially available social media tools analyze indicators such as sentiment and engagement, none offer stance detection (Ng and Carley, 2022; Rogan et al., 2022).

Large Language Models for Stance Detection: The application of LLMs, particularly ChatGPT, in stance detection has shown mixed results. Zhang et al. (2022) demonstrated improved performance on the SemEval2016 benchmark with instruction-based prompts, while other studies found moderate success (Ziems et al., 2023; Liyanage et al., 2023; Zhang et al., 2023b). However, Aiyappa et al. (2023) highlighted concerns about potential data contamination in ChatGPT's training data, and Mets et al. (2023) observed competitive but suboptimal performance in zero-shot stance classification compared to supervised models. LLMs hold promise in stance detection but may require fine-tuning or prompt engineering for optimal results.

Weak Supervision for Dataset Creation: Weak supervision has become a powerful method for efficiently generating large labeled datasets by combining heuristic rules, distant supervision, and domain expertise to produce high-quality labels from noisy sources (Ratner et al., 2019, 2016). This approach is particularly valuable for stance detection, where labeled data is scarce or costly. Recent studies have explored combining weak supervision with LLMs for improved labeling. For instance, Smith et al. (2024) and Huang et al. (2024) used LLMs to generate weak labels that were subsequently refined through weak supervision, resulting in higher labeling accuracy compared to zero-shot LLM outputs alone.

III. Dataset Creation

This dataset was collected to examine how YouTube users respond to the US Army's marketing material, thereby gauging their support or opposition toward the military. Each comment is labeled for two stances: (a) the stance of the comment toward its source video, and (b) the stance of the comment toward the US Army. Three stance labels are possible: "supports," "against," and "neutral." The following sections detail the data collection and labeling procedures.

A. Data Collection

We collected comments on videos from the US Army's official YouTube channel (username: @usarmy[3]) using the YouTube Data API[4]. This process retrieved both comments and replies to comments. Data collection took place from October 2–5, 2023, and April 2–9, 2024, yielding a total of 216,385 comments across 1,082 videos. For each comment, we collected the following information:

- **id**: the unique ID of a comment
- **comment**: the text of the comment
- **author**: a unique ID for the comment's author
- **like_count**: the number of likes the comment received
- **published_at**: the date and time of the comment's publication
- **conversation_id**: a unique ID of the conversation to which the comment belongs; this is also the ID of the root comment
- **video_id**: the unique ID of the video associated with the comment
- **name**: the title of the video associated with the comment
- **description**: a text description of the video associated with the comment
- **timestamp**: the date and time of the publication of the video the comment references

B. Data Annotation Methodology

Instead of relying on human annotators for stance labeling, we applied weak supervision Ratner et al. (2019, 2016). This approach produces a final, more accurate label by aggregating multiple lower-quality labels, known as *weak labels*, which may capture various aspects of the concept being labeled. Such techniques have been used in labeling happiness through cognitive appraisal dimensions Liu and Jaidka (2023) and identifying negotiation strategies based on reasoning and friendliness Jaidka et al. (2023). Our method used a combination of *weak labels* derived from subject matter expertise, domain knowledge, machine learning models, and recent advances in LLMs for stance detection. These labels were then combined

[3]https://www.youtube.com/usarmy

[4]https://developers.google.com/youtube/v3/docs/comments/list

through a weak supervision model to produce the final labels. Thus, our approach incorporates both specialized labeling functions, similar to those in Huang et al. (2024), and direct LLM-based labels as in Smith et al. (2024). Figure 1 illustrates the full data annotation methodology workflow. Key elements of this workflow are detailed in the following sections.

C. Annotating Weak Labels

To create the weak labels, we leveraged dataset-specific traits, domain expertise, and LLMs. In this dataset, for instance, the presence of hate speech or sarcasm frequently indicates opposition to the video and, by extension, to the US Army as the video uploader. Therefore, the presence of these weak signals can be combined to infer the final label. Importantly, each weak label does not need to produce a label for every comment in the dataset; some weak labels may not apply to certain comments. In these cases, they return an "abstain" response, which is distinct from the possible stance labels and is not used in computing the final stance labels.

1) Hate Speech: Hate speech is defined as language intended to incite violence or hatred, typically targeting specific groups such as ethnicities or religions Fortuna and Nunes (2018). In this dataset, we observed that many comments opposing both the videos and the US Army in general contained elements of hate speech, such as antisemitic or misogynistic content. A variety of machine learning algorithms, from logistic regression to neural networks, have been developed for automatic hate speech detection Fortuna and Nunes (2018). For this weak label, we used the hate speech classification model from Kralj Novak et al. (2022) [5]. This model was used as-is, without fine-tuning. For each input text, this model returns one of three different types of "hate" labels or an "acceptable" label. We interpreted any "hate" response as an "against" stance label, and any other response as "abstain."

2) Sarcasm: Sarcasm is a rhetorical device used to convey intention indirectly Chaudhari and Chandankhede (2017). Similar to hate speech, we found that commenters on these YouTube videos often used sarcasm to criticize either the video or the entity behind it. We used a multilingual sarcasm detector[6] that trains a BERT-based sarcasm classifier on newspaper text in English, Dutch, and Italian. For each input text, this model returns either a "sarcasm" or "no_sarcasm" label. We interpreted a "sarcasm" response as an "against" stance label, and any other response as "abstain."

3) Sentiment: Sentiment analysis provides insight into the positive or negative nature of public sentiment toward a topic Ibrohim et al. (2023). Although stance is distinct from sentiment, the two often overlap; negative sentiment can often indicate opposition to a target, while positive sentiment may suggest support. We used a multilingual-cased sentiment analysis model[7] that performs sentiment analysis using a teacher-student architecture with the DistillBERT model. For each input text, this model returns a score between 0 and 1 for three possible sentiment labels: "positive," "negative," and "neutral." We interpreted a "positive" sentiment score greater than 0.75 as a "supports" stance, a "negative" sentiment score greater than 0.75 as an "against" stance, and any other value as "abstain."

4) Use of Keywords: We manually curated a set of keywords and phrases that would indicate either of the non-neutral stance labels when present. For each comment, we checked for the presence of any of these keywords and assigned the corresponding stance to this weak label. For example, keywords and phrases like "hooah," "awesome," and "god bless" were associated with a "supports" stance, while terms like "woke," "murder," and "propaganda" were linked to an "against" stance. The full list of these words is available in our released code[8].

5) Stance from Large Language Models: Based on recent work demonstrating the effectiveness of LLMs for stance classification tasks (Ng et al., 2024), we also labeled stances directly using LLMs. For this, we opted to use relatively small LLM models that can run on consumer hardware, enabling scalability for large datasets. This approach was also cost-effective compared to closed-source models like GPT-4 or Claude, which quickly become cost-prohibitive for large-scale labeling Huang et al. (2024). Specifically, we used three different LLMs: (a) Flan-UL2 (Tay et al., 2022), an encoder-decoder architecture; (b) Mistral-7B (Jiang et al., 2023), a decoder-only architecture; and (c) Llama-3-8B Dubey et al. (2024), another decoder-only model. We used the Flan-UL2 and Llama-3 models solely in a zero-shot setting (i.e., without any fine-tuning) (Ziems et al., 2023).

For the Mistral-7B model, we used it both in a zero-shot setting and fine-tuned it on a collection of benchmark stance datasets, including Semeval2016 (Mohammad et al., 2016), Election2016 (Sobhani et al., 2017), srq (Villa-Cox et al., 2020), wtwt (Conforti et al., 2020), phemerumors (Kochkina et al., 2018), and covid-lies (Hossain et al., 2020). Research in (Ng and Carley, 2022) found that including more benchmark datasets in training generally improves model generalizability, so we trained on all these datasets. For the Mistral LLM training, we used a LoRA adapter (Hu et al., 2021) with parameters $\alpha = 1$, $r = 16$, $dropout = 0.1$, and no bias term. The model was trained for two epochs using an AdamW optimizer (Loshchilov and Hutter, 2017), a learning rate of $1e-4$, and a linear learning rate scheduler with 50 warm-up steps. In fine-tuning experiments, we observed that the adapters converged quickly, with validation loss stabilizing within two epochs.

To structure the data for training, we used a prompt that included context about stance classification, as commonly done in zero-shot prompting for stance tasks. The training prompt was as follows:

> "'Stance classification is the task of determining the expressed or implied opinion, or stance, of a statement toward a certain, specified target. Analyze

[5]https://huggingface.co/IMSyPP/hate_speech_en

[6]helinivan/multilingual-sarcasm-detector

[7]lxyuan/distilbert-base-multilingual-cased-sentiments-student

[8]https://doi.org/10.5281/zenodo.10493803

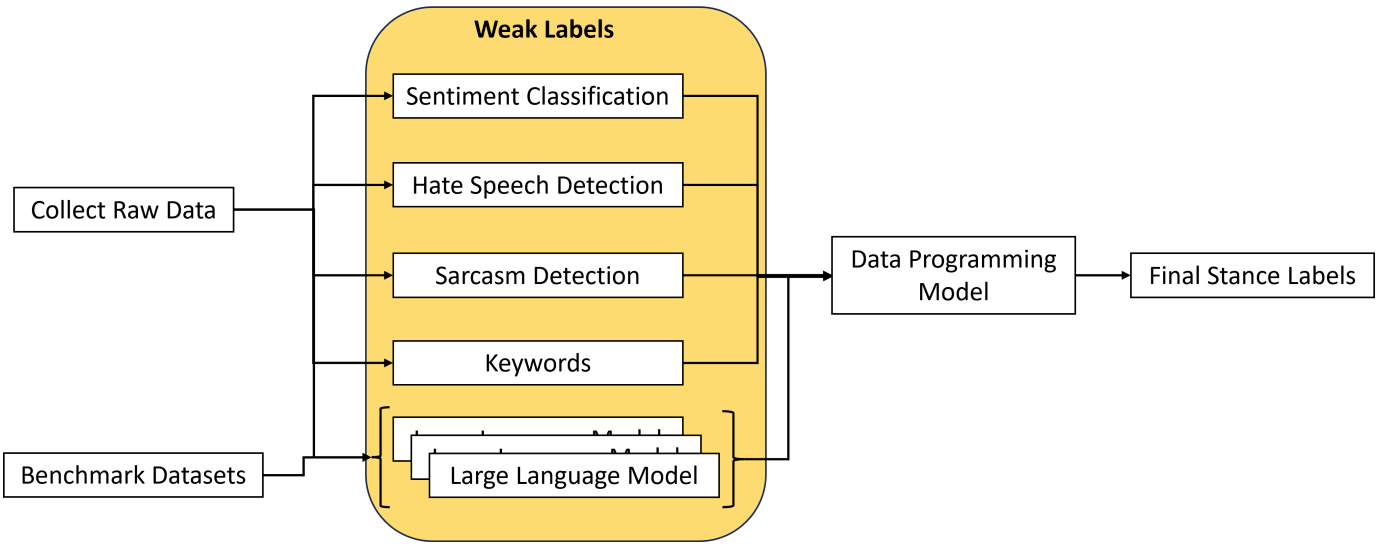


Fig. 1: Pipeline of stance labeling methodology. Stances are labeled by various heuristic means and LLMs, and then combined together with weak supervision to produce the final labels

> the following social media statement and determine its stance towards the provided [dataset entity]. Respond with a single word: "for", "against","neutral".
> [dataset entity]: {entity}
> statement: {statement to classify}
> stance: "'

The [dataset entity] was varied depending on the benchmark dataset (e.g., for Election2016, it was 'politician,' while for phemerumors it was 'rumor'). For the srq dataset, we modified the prompt to incorporate the post that the reply responds to, as that dataset labels the stance of replies toward preceding posts rather than a specific topic or person. The loss was computed based on whether the model outputted only the tokens for the correct stance. Training took place on a local computer running Ubuntu Linux with a 96-core processor and three A6000 GPUs. This setup produced four distinct LLMs for labeling: one off-the-shelf UL2 model, one off-the-shelf Llama-3 model, and two Mistral models.

Additionally, we fine-tuned a small BERT model, as in Ng and Carley (2022), on the benchmark datasets described above and used it to label stances as well.

Analyzing stances with LLMs: To analyze the comments for stance, we used two different prompts: one that included only the comment and another that included both the comment and any previous comment it was replying to. The labeling prompt without incorporating the reply structure was:

> "'Analyze the following YouTube comment to a video posted by the U.S. Army named "{title}" and determine its stance towards the provided entity. Respond with a single word: "for", "against", "neutral", or "unrelated". Only return the stance as a single word, and no other text.
> entity: {entity}
> comment: {comment to classify}
> stance:"'

while the labeling prompt that did incorporate the reply structure is:

> "'Analyze the following YouTube comment to a video posted by the U.S. Army named "{title}" and determine its stance towards the provided entity. If the statement is a reply to another YouTube comment, that YouTube comment is listed below. Respond with a single word: "for", "against", "neutral", or "unrelated". Only return the stance as a single word, and no other text.
> entity: {entity}
> previous comment: {previous_comment}
> comment: {comment to classify}
> stance:"'

We also experimented with a question-answering phrasing of the task, but found the "analyze" phrasing generally performed slightly better on benchmark datasets. Each comment was labeled for stance seven times: once with and once without the reply prompt for the UL2 model, Llama-3 model, and both Mistral models, and once without the reply prompt using the stance-tuned Mistral model. Since we labeled each comment for its stance towards both the US Army and the specific video, this resulted in 14 LLM labeling runs across the entire dataset.

To combine the LLM-derived labels, we ensembled the LLM labels using a majority vote. Specifically, we assigned a stance if more than half (i.e., 4 or more of the LLM labelers) agreed on a label for a comment; otherwise, we assigned the label as "abstain." While this slightly reduced coverage, it improved the overall quality of the LLM-derived labels.

6) Use of Analogous Stances: To label the stance of each comment toward its respective video, we also used an analogous stance as a weak label. Specifically, we used the stance

of the comment toward the U.S. Army as a proxy label for its stance toward the video. Generally, if a comment was against the U.S. Army, it was also against a video posted by the U.S. Army. Prior research suggests that leveraging complementary stances can improve performance in stance detection tasks Sobhani et al. (2017); Zhang et al. (2023b). Consequently, we used the final stance toward the U.S. Army, derived through weak supervision on the previously described weak labels, as a weak label for the stance toward the video. It is important to note that this relationship does not apply in reverse; we found several examples where commenters expressed support for the U.S. Army while criticizing a particular video.

D. Final Stance Determination

With the weak labels for the comments established, we applied a weak supervision methodology using the Snorkel Python package[9]. Snorkel utilizes a unified generative model that accounts for partial coverage of the weak labels and the correlations between them to produce final labels. For further details on this methodology, including training and label generation, we refer the reader to Ratner et al. (2016) and Ratner et al. (2019). The final stance labels can take one of three values: "supports," "against," or "neutral" for both the stance toward the video and the stance toward the U.S. Army.

E. Human Evaluation

To assess the quality of the machine-generated labels, we conducted a human evaluation. Given the impracticality of labeling the entire dataset manually, we randomly selected a subset of 1,000 comments, which were labeled by four annotators. We evaluated the inter-annotator agreement and consolidated the human-provided labels through majority voting. Using this subset, we then assessed the quality of the machine-produced labels.

IV. Data Properties and Analysis

This section describes some exploratory data analysis performed on this dataset. The dataset consists of 1082 videos published from 1 June 2010 through 1 April 2024 on the US Army's official YouTube Channel. From these videos, there are a total of 215,509 comments.

A. Dataset Properties

Distribution of videos over time: In general, the channel has published an increasing number of videos over time, with occasional periods of lower activity. Figure 3 illustrates the distribution of all videos on the channel over time, while Table II provides summary statistics for the comments associated with each video. Most videos in the dataset have only a single comment, with the number of comments per video being highly skewed; some videos have a substantial number of comments (the most commented-on video, "Soldier wakes up and chooses 'HOOAH!'," contains 9.4% of all comments by itself). In terms of comment timing, the vast majority have been posted recently in the channel's timeline, particularly since 2022.

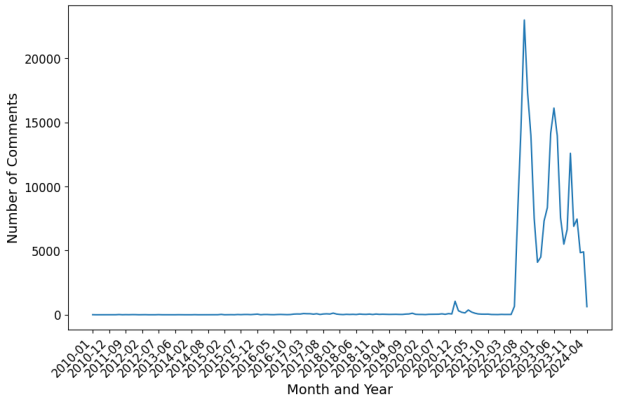


Fig. 2: Distribution of comments over time

As shown in Figure 2, the timeline of comments reveals notable spikes in August, September, and October of 2022, and again in May and June of 2023. For the first spike, the most commented-on video was "Soldier wakes up and chooses 'HOOAH!'," whereas, in the Spring 2023 spike, "Words of wisdom from a #WWII veteran" received the most comments. The commentary on the first video reflects a divisive split in opinions on the Army's role and actions, while the comments on the second video express nostalgia for the Army's role in World War II alongside concerns about its current state. Examining the comments in more detail, we find that most are relatively short, in line with typical social media posts, though a few extend to over 1,000 words. Table I presents the word count statistics for the comments.

Word-level Statistics of Comments	Value
Mean	16.66
Standard Deviation	33.13
Minimum	1.00
25th Percentile	4.00
50th Percentile (Median)	9.00
75th Percentile	18.00
Maximum	1851.00

TABLE I: Word-level Statistics of the Comments

Statistics of comments per video	Value
Mean	199.17
Standard Deviation	883.41
Minimum	1
25th Percentile	7
50th Percentile (Median)	37
75th Percentile	86
Maximum	18999
Mode	1

TABLE II: Per Video Comment Statistics

B. Stance Label Properties

In examining the stance labels generated through our methodology, we find a slight skew toward the "against" stance

[9]https://snorkel.readthedocs.io/en/master/index.html

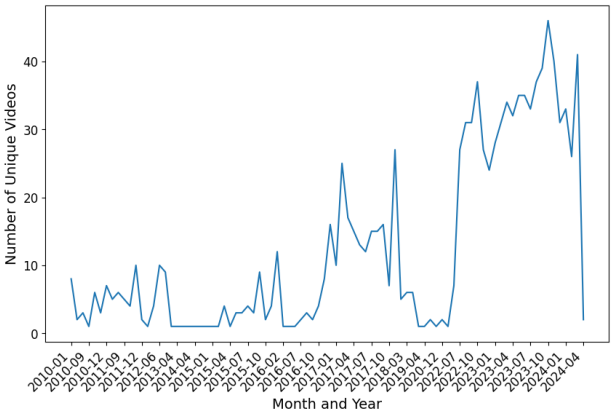


Fig. 3: Distribution of videos published over time

for both the U.S. Army and the videos posted on the channel. Table III summarizes the counts for each stance label.

From the labeling, we observe some degree of label skew. First, there is a noticeable skew toward the "neutral" label class, consistent with patterns in other stance-labeled datasets (e.g., Hossain et al. (2020) and Kochkina et al. (2018)). Additionally, we observe a slight skew toward "against" stances over "supportive" stances. Given that these are comments typically intended to support a brand (i.e., the U.S. Army), the prevalence of negative stances toward the U.S. Army and its videos is noteworthy. Lastly, we note that around 1% of the data remains unlabeled for each stance target due to limitations in the weak labeling process.

Stance	U.S. Army	Video
abstain	2373 (1.10%)	1564 (0.73%)
against	49817 (23.12%)	63222 (29.34%)
neutral	122028 (56.61%)	102479 (47.56%)
supports	41291 (19.16%)	48244 (22.37%)

TABLE III: Label counts (and percentages of the total) for each stance target

Analyzing the distribution of stance labels over time, we find that all three stance classes tend to trend together. **??** displays the distribution of stance classes toward the U.S. Army across all comments and towards their respective videos, across all videos. Stance distributions toward the videos follow a similar pattern. Notably, the two prominent peaks of comment activity differ in stance composition: in the first peak (September 2022), "against" labels are more prevalent than "supports," while in the second peak (May-June 2023), "supports" labels outnumber "against." Overall, this suggests a moderate balance among stances toward the videos, with shifts over time and between individual videos.

To further understand stance differences, we examined the videos with the most positive, neutral, and negative receptions. Among videos with over 100 comments, "Honoring Our Fallen Heroes - Taps" had the highest percentage of supportive comments (61.5%), while "Sergeant Major of the Army Grinston addresses the report of the Fort Hood Independent Review" had the most "against" comments (55.9%). The most neutrally received video, "Army #PopQuiz - What training is this?", received 81.6% neutral comments, with many simply answering the question posed in the video.

Finally, since this dataset includes dual stances for each comment, we can examine the relationship between stances toward the U.S. Army and the videos. Figure 5 displays a confusion matrix for these two stance targets. Unsurprisingly, the stances closely resemble each other; however, there are significant instances where comments support the video but are neutral toward the Army, or are critical of the video but supportive of the Army. Manual inspection reveals that the first category often contains positive remarks about the video alone (e.g., "that video was awesome!"), while the second group expresses frustration with the video while generally supporting the Army (e.g., "What is this bs? The Army is better than this…"). While the two stance labels are closely related in this dataset, the observed divergences underscore the complexity of stance in this domain.

C. Comparison with Human Evaluation

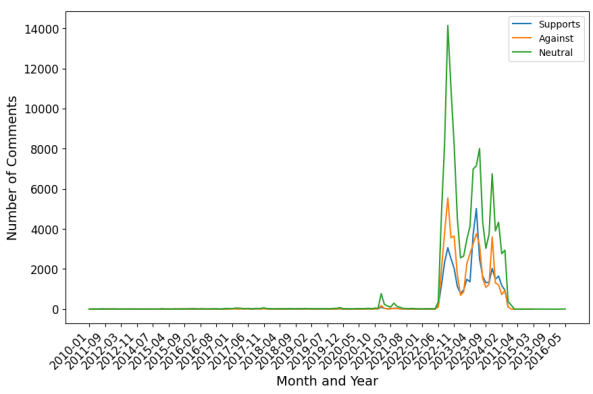
Finally, we assessed the performance of machine-generated stance labels using state-of-the-art techniques. To do this, we created a validation dataset, as described earlier, and had four human labelers annotate the dataset. For the stance toward the U.S. Army, the average Cohen's Kappa McHugh (2012) across all human labelers was 0.863, and for the stance toward the videos, it was 0.797. These high Kappa values indicate strong agreement among the human labelers regarding the stance of the comments (note that we have also released this labeled validation dataset).

When comparing the machine-generated labels from the weak supervision process, we found moderately successful performance. For the stance toward the U.S. Army, the accuracy and Micro F1-scores were 0.68, while for the stance toward the videos, the accuracy and Micro F1-scores were 0.639. When comparing the performance of various LLMs for the labeling task (see Table IV), we observed that the weak supervision-generated labels performed better than the best-performing LLM-generated labels, particularly for the stance toward the videos.

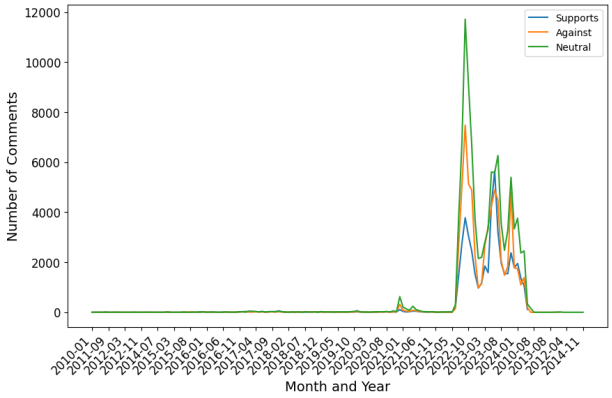
Labeling Model	U.S. Army	Video
mistral zero-shot with comments	66.90%	56.50%
llama-3 zero-shot with comments	41.40%	50.90%
mistral zero-shot	64.40%	55.10%
llama-3 zero-shot	41.20%	40.20%
ul2 zero-shot with comments	53.00%	45.00%
ul2 zero-shot	49.00%	45.80%
mistral adapter	49.70%	26.50%
small model	52.50%	32.30%

TABLE IV: Micro-F1 scores for U.S. Army and Video stances, by different models and prompts, on the validation dataset.

From the analysis of the LLM labelers, we observe significant differences in performance between models and prompting schemes. The Mistral LLM Jiang et al. (2023) and the prompt scheme incorporating comments with replies generally performed better.



(a) Stance toward U.S. Army



(b) Stance toward the videos

Fig. 4: Distribution of stance labels over time for comments on the U.S. Army's official channel toward the Army and the videos posted on the channel.

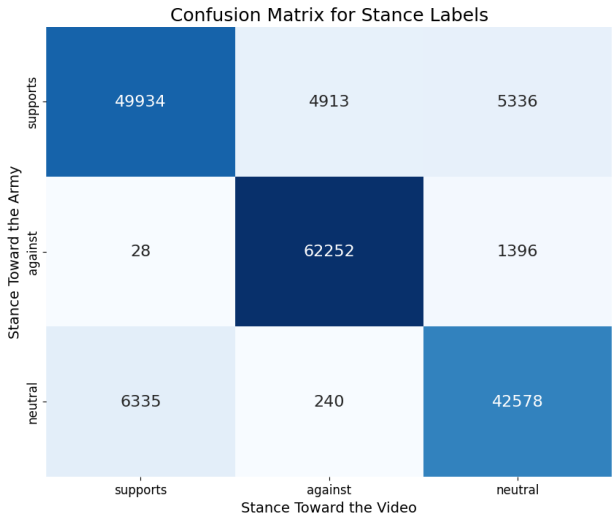


Fig. 5: Confusion Matrix of Stance Toward the U.S. Army and Stance Toward the Video

To better understand these results, we plotted the confusion matrix comparing human-derived and machine-derived labels, shown in Figure 6. From these matrices, we observe that for valenced stances, machine-generated stances generally align with human-annotated stances. The majority of the accuracy loss in machine-generated labels stems from misclassifying neutrally-stanced comments. Upon closer inspection, we found that some of these misclassified examples involved highly positive-sounding comments that were irrelevant to either the video or the Army, as well as comments where sarcasm or coded language (e.g., "goy", "zog") was missed by the LLMs.

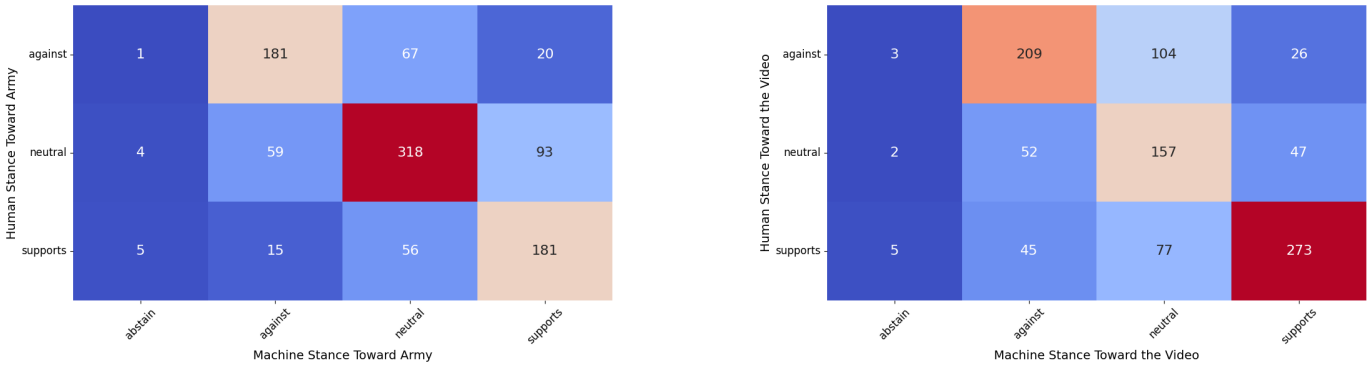
V. Discussion

Public opinion toward a country's military is crucial, as it signals the level of trust in the government and the willingness of citizens to join the military. A positive opinion facilitates military recruitment and support, especially in times of crisis. Social media engagement plays a pivotal role in shaping this opinion. This paper introduces a dataset that supports understanding and associated analyses, comprising comments on the U.S. Army's official YouTube videos, annotated for their stance toward the Army and the video itself.

To construct this dataset, we employed weak supervision, which leverages *weak labels* for automated labeling. This paradigm provides a transition from direct human annotation to machine-assisted annotation, allowing scalability to larger datasets. For our specific dataset, we utilized subject matter expertise, domain knowledge, and recent advancements in using LLMs for stance determination to create weak labels. Specifically, we used weak labels such as the presence of hate speech, sarcasm, and sentiment as indicators for certain stances. This transformation converted the challenging, ambiguously defined problem of determining and labeling stance into a more well-defined problem for specific aspects or stances. These could then be aggregated through weak supervision to produce comprehensive stance labeling. For instance, the presence of hate speech indicates a comment against the video and, by extension, against the user who uploaded it (e.g., the Army).

We further enhanced these weak labels with contributions from LLMs. Currently, the effectiveness of LLMs in tasks like stance labeling remains uncertain. While recent research shows promise for using LLMs, especially when combined with prompt engineering, for stance detection and labeling, the generalizability of these methods and their compatibility with all LLMs is unclear. Through weak supervision, we treat these LLM-derived labels as noisy yet valuable, enhancing the creation of a stance-labeled dataset. The combination of dataset-specific weak signals and the linguistic understanding power of LLMs produces high-quality labels for stance in a more scalable manner compared to purely manual or machine-generated labels.

In this work, we combined existing state-of-the-art methods for automated labeling of a dataset on the difficult task of stance labeling. We found that these approaches were able to



(a) Confusion matrix for stance toward the U.S. Army.

(b) Confusion matrix for stance toward the video.

Fig. 6: Confusion matrices between the machine-derived labels and the human-derived labels for the validation set.

produce reasonable labels relative to a human-labeled validation dataset, but there is still room for improvement in the development of automated labeling techniques, especially for challenging labeling tasks like stance. This opens avenues for further exploration of improving stance classification methods, especially for out-of-domain tasks, to provide more accurate labels toward an entity and therefore more accurate downstream insights.

Potential Applications: The dataset proposed in this work has several potential uses. First, it presents a versatile resource for research on the analysis of stance in YouTube video comments. Beyond its primary application in training and evaluating new stance detection techniques and models, the dataset introduces distinctive traits not present in current benchmarks. Notably, it originates from YouTube, offering a departure from the prevalent X (formerly Twitter)-based datasets. Furthermore, the dataset features multiple stance targets and exhibits a substantial degree of stance-taking behavior and controversial content (i.e., a high proportion of "supports" and "against" stances). The dataset provides sequential information about comments, enabling investigations into the dynamics of stance relative to the stance expressed in a previous comment.

Given the dataset's proximity to a politically and culturally divisive topic — the military — it encompasses a spectrum of online behaviors, including the propagation of conspiracy theories, trolling, hate speech, and misinformation. As such, it serves as a valuable resource for studies on misinformation and disinformation.

Finally, the dataset's temporal collection allows researchers to explore dynamic aspects of stance-taking behavior over time. This temporal dimension provides insights into the evolution of discussions and sentiments within the context of the U.S. Army's YouTube channel. In summary, the dataset offers a rich foundation for various studies, ranging from advancing stance detection methodologies to delving into the nuanced dynamics of online interactions related to a prominent and controversial subject.

Dataset Availability: We collected data using the YouTube API and did not perform additional scraping. We only collected public videos and comments, and made no attempt to collect private information. We anonymized the author IDs and did not perform any author-level analysis. The code used to construct the dataset and the dataset itself are provided at https://doi.org/10.5281/zenodo.10493803.

Limitations and Future Work: Our dataset is limited to comments from one channel and the comments we could collect (not all videos posted to the channel allowed comments). We hope to expand our dataset to include comments on videos posted by armies around the world, to evaluate public opinion toward the military globally. Additionally, our dataset is limited to relatively small LLMs due to computational constraints. In the future, we aim to expand the dataset annotation to include different types of LLMs with varying prompting schemes to better understand the stance annotation capabilities of LLMs.

VI. Conclusion

Understanding the stance expressed in YouTube comments toward an entity is essential for gaining insights into public opinion regarding the authors of the videos. In this work, we introduce the DIVERSE dataset, which captures stances toward the U.S. Army. This dataset is formulated from comments gathered from the U.S. Army's YouTube videos. We constructed the final stance variable utilizing weak labels, including the identification of hate speech and sarcasm, and incorporated stance inferences from Large Language Models. By employing weak supervision, we amalgamated these diverse labels to generate the final stance annotations. We anticipate that this dataset will prove invaluable to researchers investigating public sentiment and stances toward military entities.

Acknowledgements

The research for this paper was supported in part by the Center for Informed Democracy and Social-Cybersecurity (IDeaS) at Carnegie Mellon University. The views and conclusions expressed are those of the authors and should not be interpreted as representing the official policies, either

expressed or implied, of the Department of Defense, the U.S. Army, or the U.S. Government.

REFERENCES

E. Brown, "The ghost of gwot haunting the military recruiting crisis," *The Modern War Institute*, 2023. [Online]. Available: https://mwi.westpoint.edu/the-ghost-of-gwot-haunting-the-military-recruiting-crisis/

A. AlDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, 2021.

L. H. X. Ng and I. J. Cruickshank, "Recruitment promotion via twitter: a network-centric approach of analyzing community engagement using social identity," *Digital Government: Research and Practice*, vol. 4, no. 4, pp. 1–17, 2023.

E. Allaway and K. McKeown, "Zero-shot stance detection: Paradigms and challenges," *Frontiers in Artificial Intelligence*, vol. 5, p. 1070429, 2023.

J. W. Du Bois, "The stance triangle," *Stancetaking in discourse: Subjectivity, evaluation, interaction*, vol. 164, no. 3, pp. 139–182, 2007.

P. Sobhani, D. Inkpen, and X. Zhu, "A dataset for multi-target stance detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 551–557.

C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier, "Will-they-won't-they: A very large dataset for stance detection on twitter," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1715–1724.

S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 31–41.

T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, "COVIDLies: Detecting COVID-19 misinformation on social media," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3402–3413.

R. Villa-Cox, S. Kumar, M. Babcock, and K. M. Carley, "Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations," *arXiv preprint arXiv:2006.00691*, 2020.

L. H. X. Ng and K. M. Carley, "Is my stance the same as your stance? a cross validation study of stance detection datasets," *Information Processing & Management*, vol. 59, no. 6, p. 103070, 2022.

A. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré, "Training complex models with multi-task weak supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4763–4771.

T.-H. Huang, C. Cao, V. Bhargava, and F. Sala, "The alchemist: Automated labeling 500x cheaper than llm data annotators," *arXiv preprint arXiv:2407.11004*, 2024.

R. Smith, J. A. Fries, B. Hancock, and S. H. Bach, "Language models in the loop: Incorporating prompting into weak supervision," *ACM/JMS Journal of Data Science*, vol. 1, no. 2, pp. 1–30, 2024.

V. Qazvinian, E. Rosengren, D. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 1589–1599.

H. Elfardy and M. Diab, "Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 434–439.

I. Augenstein, A. Vlachos, and K. Bontcheva, "Usfd at semeval-2016 task 6: Any-target stance detection on twitter with autoencoders," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 389–393.

W. Fang, M. Nadeem, M. Mohtarami, and J. Glass, "Neural multi-task learning for stance prediction," in *Proceedings of the second workshop on fact extraction and verification (FEVER)*, 2019, pp. 13–19.

U. A. Siddiqua, A. N. Chy, and M. Aono, "Tweet stance detection using an attention based neural ensemble model," in *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, 2019, pp. 1868–1873.

D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, 2020.

B. Zhang, D. Ding, L. Jing, and H. Huang, "A logically consistent chain-of-thought approach for stance detection," *arXiv preprint arXiv:2312.16054*, 2023.

J. Gatto, O. Sharif, and S. M. Preum, "Chain-of-thought embeddings for stance detection on social media," *arXiv preprint arXiv:2310.19750*, 2023.

T. Marcoux, N. Agarwal, R. Erol, A. Obadimu, and M. N. Hussain, "Analyzing cyber influence campaigns on youtube using youtubetracker," *Big Data and Social Media Analytics: Trending Applications*, pp. 101–111, 2021.

M. N. Hussain, S. Tokdemir, N. Agarwal, and S. Al-Khateeb, "Analyzing disinformation and crowd manipulation tactics on youtube," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 1092–1095.

S. Y. Liaw, F. Huang, F. Benevenuto, H. Kwak, and J. An, "Younicon: Youtube's community of conspiracy videos," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 1102–1111.

A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal, "Identifying toxicity within youtube video comment," in *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*. Springer, 2019, pp. 214–223.

C. Rogan, S. Davic, S. Gatsios, and I. Lachow, "A review of commercial capabilities that focus on online influence," MITRE, McLean, Vriginia, Technical Report MTR220415, 2022.

B. Zhang, D. Ding, and L. Jing, "How would stance detection techniques evolve after the launch of chatgpt?" *arXiv preprint arXiv:2212.14548*, 2022.

C. Ziems, O. Shaikh, Z. Zhang, W. Held, J. Chen, and D. Yang, "Can large language models transform computational social science?" *Computational Linguistics*, pp. 1–53, 2023.

C. Liyanage, R. Gokani, and V. Mago, "Gpt-4 as a twitter data annotator: Unraveling its performance on a stance classification task," *Authorea Preprints*, 2023.

B. Zhang, X. Fu, D. Ding, H. Huang, Y. Li, and L. Jing, "Investigating chain-of-thought with chatgpt for stance detection on social media," *arXiv preprint arXiv:2304.03087*, 2023.

R. Aiyappa, J. An, H. Kwak, and Y.-Y. Ahn, "Can we trust the evaluation on chatgpt?" *arXiv preprint arXiv:2303.12767*, 2023.

M. Mets, A. Karjus, I. Ibrus, and M. Schich, "Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media," *arXiv preprint arXiv:2305.13047*, 2023.

A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: Creating large training sets, quickly," *Advances in neural information processing systems*, vol. 29, 2016.

X. Liu and K. Jaidka, "I am psyam: Modeling happiness with cognitive appraisal dimensions," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1192–1210.

K. Jaidka, H. Ahuja, and L. Ng, "It takes two to negotiate: Modeling social exchange in online multiplayer games," *arXiv preprint arXiv:2311.08666*, 2023.

P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, jul 2018. [Online]. Available: https://doi.org/10.1145/3232676

P. Kralj Novak, T. Scantamburlo, A. Pelicon, M. Cinelli, I. Mozetič, and F. Zollo, "Handling disagreement in hate speech modelling," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2022, pp. 681–695.

P. Chaudhari and C. Chandankhede, "Literature survey of sarcasm detection," in *2017 International conference on wireless communications, signal processing and networking (WiSPNET)*. IEEE, 2017, pp. 2041–2046.

M. O. Ibrohim, C. Bosco, and V. Basile, "Sentiment analysis for the natural environment: A systematic review," *ACM Comput. Surv.*, vol. 56, no. 4, nov 2023. [Online]. Available: https://doi.org/10.1145/3604605

L. H. X. Ng, I. Cruickshank, and R. K.-W. Lee, "Examining the influence of political bias on large language model performance in stance classification," *arXiv preprint arXiv:2407.17688*, 2024.

Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, D. Bahri, T. Schuster, H. S. Zheng, N. Houlsby, and D. Metzler, "Unifying language learning paradigms," *arXiv preprint arXiv:2205.05131*, 2022.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.