Annotating Twitter Data from Vulnerable Populations: Evaluating Disagreement Between Domain Experts and Graduate Student Annotators

Desmond U. Patton
Columbia University
dp2787@columbia.edu

Philipp Blandfort
DFKI
philipp.blandfort@dfki.de

William R. Frey
Columbia University
wf2220@columbia.edu

Michael B. Gaskell
Columbia University
mbg2174@columbia.edu

Svebor Karaman
Columbia University
svebor.karaman@columbia.edu

Abstract

Researchers in computer science have spent considerable time developing methods to increase the accuracy and richness of annotations. However, there is a dearth in research that examines the positionality of the annotator, how they are trained and what we can learn from disagreements between different groups of annotators. In this study, we use qualitative analysis, statistical and computational methods to compare annotations between Chicago-based domain experts and graduate students who annotated a total of 1,851 tweets with images that are a part of a larger corpora associated with the Chicago Gang Intervention Study, which aims to develop a computational system that detects aggression and loss among gang-involved youth in Chicago. We found evidence to support the study of disagreement between annotators and underscore the need for domain expertise when reviewing Twitter data from vulnerable populations. Implications for annotation and content moderation are discussed.

1. Introduction

Annotation is the process of providing metadata (e.g. deeper meaning, context, nuance) through the act of labeling language or other contents such as images or videos. Machine learning and natural language research has long relied on the robust annotation of social media data to examine and predict myriad human phenomenon [10, 12, 14]. In the context of machine learning, the annotation process typically involves assigning categories to items, which are then used to build computational models for detecting these categories [1, 9]. With an understanding that language is highly subjective, researchers in computer science have spent considerable time developing new methods to increase the richness of annotation [10] and combine annotations stemming from multiple annotators [18, 21, 25] based on estimated reliabilities [14, 19]. Most of these efforts have focused on inter-annotator reliability, improving accuracy across annotators and reducing disagreement regarding how to interpret data [10], often without analyzing causes of disagreement [14, 18, 19, 21]. Furthermore, these methods assume that for each given item there is one "correct" label. However, when human annotators disagree when choosing a different label for the same post, one must consider if there actually is a single correct answer. In addition, if an annotator holds more contextual knowledge than another, should some patterns of disagreements be weighed more heavily than others [19]? To extend this idea, we build on the work of Brock [6] and Roberts [20] who underscore the importance of centering the perspectives, viewpoints, and epistemologies of vulnerable and marginalized communities when analyzing social media data.

On the other hand, there is a gap in research which examines the positionality who annotates the data (e.g. demographics, expertise, experience), how they are trained and the extent to which those characteristics impact how data is labeled and interpreted. A deeper focus on annotation is particularly important when analyzing data from vulnerable and marginalized populations on social media. Symbolic interactionism theory suggests that the ways in which we derive meaning is in response to an interpretive process based in our social interaction with others [5]. That is to say, the meaning of social media posts from African American youth in Chicago and how they should be interpreted is rooted in a nuanced understanding of the everyday activities, experiences, language and shared culture. As such, the expertise and training of the annotators are important when observing local concepts, gestures, innuendo, and other psycho-social scripts and

URI: https://hdl.handle.net/10125/59653
ISBN: 978-0-9981331-2-6

behaviors embedded in text and images on social media. For example, in her book "It's Complicated", danah boyd describes a young African American male high school student who loses his spot at Yale University because of images on his Facebook profile that were interpreted as being connected to gang involvement. Misinterpreting nuances in language, culture, and context can have detrimental consequences that lead to criminalization and further stigmatization of marginalized groups [7, 16]. Determining when and if something is inappropriate is highly subjective and at the whim of annotators and content moderators who may have no familiarity with the language, concepts, and culture of the social media user [20].

In this paper, we present findings from the analysis of annotation within and between two groups: two African-American Chicago-based domain experts and two social work graduate students (one African American, one White) who annotated a total of 1,851 tweets with images from Twitter that are a part of a larger corpora associated with the Chicago Gang Intervention Study, which contains tweets with images from African American youth and young adults (See section 4). The broader purpose of this study is to develop a computational system that detects pathways to violence among gang-involved youth in Chicago. The paper is organized as follows. Section 2 provides a description of the annotation process. Section 3 provides a description of the methods for analysis of annotator perspectives. Section 4 introduces the case study which includes an analysis of differences in annotation within and between groups, what is revealed from those differences, and what to take from it. Section 5 describes implications from the study which include the importance of annotator training, how annotation should be monitored to identify problems, what to do with errors in annotations and how domain experts should be involved in the annotation process. Section 6 describes future directions which include other applications of our analysis methods and the implications of this work for content moderation.

2. Description of Annotation Process: The Contextual Analysis of Social Media Approach

The annotation process involves labelling tweets with respect to the psychosocial codes *aggression, loss*, and *substance use*, and contains various key components: annotators (Chicago-based domain experts and social work graduate students), social work graduate student annotator training, the Contextual Analysis of Social Media (CASM) approach [17], and a web-based visual and textual analysis system for annotation [15]. The annotation process for each group of annotators has distinctions due to their different expertise.

2.1. Chicago-based domain experts

In order to ensure an accurate and contextual understanding of the images and text embedded in our Twitter corpus, we partnered with a local violence prevention organization in Chicago to hire two individuals as domain experts. We asked the partner organization to identify individuals who had a deep understanding of the local language, concepts, gang activity, and who were active on Twitter. The partner organization identified one African American man in his early 20's, a community member, and one African American woman in her late 20's, an outreach worker for the organization. The domain experts were asked to annotate 1,851 images using the annotation system. A white postdoctoral research scientist, with a doctorate in clinical psychology and based in Chicago trained the domain experts how to use the system, validated their community expertise, and clarified the purpose of the tasks and research. The domain experts were not trained on how to define and interpret *aggression*, *loss*, and *substance use* because we intentionally center their knowledge of community, language, and experience as expertise. As such, the domain experts are educating the researchers on how to define the aforementioned classifications [8]. Domain experts annotated the entire dataset on average within 48 hours from receiving the data because of their facility with the language and content embedded in the Twitter posts.

2.2. Social work graduate students

Social work graduate student annotators were current students in a Master of Social Work program. Both students are women and in their early 20's one is African American and the other is White. They were chosen based on their professional experience in adolescent development, criminal justice, and community work with youth of color. All annotators showed and expressed an openness and willingness to learn through their prior work and participation in the SAFElab. The annotators undergo a rigorous training process involving five steps: 1) a general overview of the domain, 2) the annotator role, 3) annotation process and system tutorial, 4) deep Twitter immersion, and 5) annotation practice and feedback. The social work annotators received this specific training because they lacked the life experience that would provide them a firm understanding of the local context and language,

which could potentially lead to gross misinterpretations of the Twitter posts [16].

The training begins with an overview of the domain informed by insights from domain experts, which includes geography, relevant historical information (e.g., relationships and rivalries between gangs), and data around violence and victimhood. After the students received an overview of the domain, we outline their role as an annotator of Twitter posts. This involves describing the purpose and aims of the work and an introduction to thematic qualitative analysis of text and images. Additionally, our annotators engage with the ethical and sociopolitical aspects they will come across during annotation (e.g., privacy and protection, Twitter data from marginalized communities, implications regarding race), which includes understanding their own relation to the Twitter data and the domain [17].

Next, students are taken through CASM in our web-based annotation system, which includes instructions on accurate and efficient use of the system. CASM is a team-based contextually driven process used to qualitatively analyze and label Twitter posts for the training of computational systems. CASM involves a baseline interpretation of a Twitter post, followed by a deep analysis of various contextual features of the post, the post's author, their peer network, and community context. A thematic label is then applied to the post. These reconciled labeled posts are then provided to the data science team for computational system training. The steps of CASM are clearly outlined in the analysis system to help quickly orient each annotator.

Following the methodological and web-based system tutorial, student annotators undergo a week-long immersion on Twitter. This immersion includes passive observation of twenty Twitter users from our corpus to familiarize themselves with the dataset by going through each user's profile, posts, photos, and videos. The annotators are instructed to ethnographically observe the ways users portray themselves online through what they share, who they engage with, and how frequently they post. The Twitter immersion also involves a critical ethical discussion regarding their observation. As a group, student annotators agree to guidelines for protecting the anonymity of users, including: completion of annotations in a private setting, exclusion of users with private accounts, and separation of field notes and identifying information.

After the Twitter immersion, students attend a process meeting to share their observations with other annotators and the expert annotator (the trainer). The meeting is spent training the new annotators to consider contextual features they may be missing from their initial observations. In the second week of training, student annotators annotate 100 Twitter posts. These annotations are thoroughly reviewed by the expert annotator for any egregious mistakes and patterns of misinterpretation. Some examples of this include misunderstanding various steps of CASM, missing contextual features, and not utilizing web-based resources in the annotation process (e.g., Hipwiki). The expert annotator provides feedback and then the annotators are ready to begin the full annotation process on the official Twitter dataset.

3. Methods for Analysis of Annotator Perspectives

3.1. Qualitative

The postdoctoral research scientist conducted one interview with each domain experts that were employed by the lead author to conduct annotations. The purpose of the interview was to discuss the coding process in general and to review a subset of the annotations in detail to better understand the aspects of images that led to a specific classification. Interviews were conducted at a Chicago-based violence prevention organization in which the domain experts were either employed or affiliated. The mission of the organization is to reduce violence in Chicago by "replacing the cycle of violence using the principles, practices and teachings of nonviolence."

The social science team reviewed two main types of annotation examples. First, we selected examples where a domain expert provided a label that was unique (different from the other domain expert and from the student annotators) across four different classifications: *aggression*, *loss*, *substance use*, or no label. For both of the domain experts we selected 20 unique examples. Second, we selected an additional five examples in each of the four classifications (20 additional examples) where the domain experts agreed with each other, but the social work annotators provided a different label.

The postdoctoral research scientist then conducted separate structured interviews with each domain expert annotator for 30 to 45 minutes. The domain experts described how they interpreted and labeled the tweets. Oral consent was obtained, and both participants were paid an hourly rate for the time it took to conduct the interviews. During the interview, the annotators were asked to describe and explain their responses to 40 tweets with 20 of them overlapping between them. The postdoctoral research scientist reviewed 60 unique tweets in total, which accounts for approximately 10% of the total number of disagreements.

We analyzed the interview data using an inductive qualitative approach. The interviews were transcribed and read on once initially to create a list of preliminary codes. We then applied a codebook to the transcripts and

revised them based on a thorough read. Both transcripts were then coded by two additional authors. We resolved discrepancies through discussion until consensus was achieved. All data was analyzed by hand given the small amount of data. Emerging themes were established by reviewing the transcripts repeatedly and considering the similarities and differences of meaning across the annotators. We will discuss the findings from the interviews with domain experts in Section 4.

3.2. Statistical and computational methods

We compute several statistics for evaluating disagreements between annotators.

Code baselines. First of all, for each annotator (or group of annotators for which we merge their annotations) and code, we compute the overall proportion of positive votes. These proportions will be referred to as code baselines and can be seen as a measure for the annotator's overall tendency to label a tweet as the respective code. We compute a confidence interval for these numbers (interpreting the decisions as coming from a binomial distribution).

Annotator correlation coefficients. To obtain a general measure of how much disagreement there is in the data, for each class we compute Spearman correlation coefficients for the labels given by two annotators (or group of annotators for which annotations are merged).

Disagreement statistics. For two given annotators (or two groups of annotators for which annotations are merged) and each code, we first calculate the baseline proportion of the number of tweets with conflicting annotations to the overall number of tweets. In addition, for each (textual or visual) concept *c* we compute the same ratio but only consider tweets that contain the concept *c*. We compute confidence intervals for the baselines as well as the concept-based ratios as for the code baselines. We use statistical testing to check which concepts significantly affecting disagreement: if the confidence interval for concept *c* does not overlap with the confidence interval of the respective baseline, this means that for the chosen annotators and code, the concept *c* has a significant impact on the amount of disagreement between these annotators for this code. Such a difference indicates that the annotators might implicitly assign different code relevance to the respective concept, or, in other words, interpret the concept differently for the task at hand.

Annotator bias. To better understand the reasons for disagreement, for all concepts and codes, we compute the average direction of disagreement. To this end, we first compute differences in code labels for an individual tweet as values -1, 0 and 1 by subtracting the (binary) label of the first annotator from the label given by the second annotator. We then compute the average and confidence interval for the resulting list of non-zero values over all relevant tweets (i.e. that include the concept of interest). A baseline bias is computed over all tweets and significance is checked similar to the calculation of concept-based disagreement ratios.

Concept disagreement correlations. For each concept and code, we calculate Spearman correlation coefficients between concept presence in the tweets and disagreement in the associated annotations. This provides an additional measure for the importance of individual concepts for disagreement.

Disagreement prediction. We order tweets by annotation times and for different positions x, use the first x tweets for training logistic regression models to predict disagreement with respect to any of the codes, using textual, visual or both kinds of features as model input. All models are then evaluated on the test data which at any time consists of all tweets that have not been used for training. This method has some resemblance to the one proposed in [25] but aims at predicting disagreement instead of the label given by an individual annotator and does not assume the existence of any "true" gold label.

4. Case Study

The corpus for this study comes from the Gang Intervention and Computer Science study, an interdisciplinary project between the SAFElab at the School of Social Work and several members of the Data Science Institute at Columbia University. This project leverages publicly available data from youth and young adults who claim gang association and ties on Twitter and aims to better understand the individual, community, and societal-level factors and conditions that shape aggressive communication online and to determine potential pathways to violence using machine learning.

In order to create our Twitter corpus, we first scraped data from Gakirah Barnes. The first author has studied the Twitter communication of Gakirah Barnes since 2014. Motivations for this study included her age, race, and location, all of which the literature points to as potential risk factors for violence, victimization, and perpetration [23]. Moreover, her assumed gender, physical presentation on Twitter, status within a local Chicago gang, and mentions and subsequent conversations conducted on Twitter regarding two homicides, all made her a unique case study. Gakirah was a 17-year-old female who self-identified as a gang member and "shooter." After the murder of her close friend Tyquan Tyler, Gakirah changed her Twitter account handle to @TyquanAssassin. Gakirah was

active on Twitter, amassing over 27,000 tweets from December 2011 until her untimely death on April 11, 2014. She used the account to express a range of emotions to include her experiences with love, happiness, trauma, gang violence, and grief.

Our corpus contains 1,851 tweets from 173 unique users scraped in February 2017. Users for this corpus were selected based on their connections with Gakirah Barnes and her top 14 communicators in her Twitter network. Additional users were collected using a snowball sampling technique [2]. For each user we removed all retweets, quote tweets, and tweets without any image, and limited to 20 tweets per user as a strategy to avoid the most active users being overrepresented.

4.1. Qualitative findings

Three themes emerged from the interviews with domain experts, which accounted for the majority of differences between the domain experts and student annotators: recognizing community-level factors, people, and hand gestures.

First, domain experts were able to better recognize community-level factors like places or context. For example, a domain expert identified a handmade card in one of the images. She explained that this type of card was made in and sent from prison. This contextual clue influenced a decision to categorize the photo as *loss*. In another example, a home was featured prominently in a Twitter photo, which had a line of people waiting in front of the house. Both domain experts suggest that this photo presented a house used to distribute illicit drugs. Second, domain experts recognized certain individuals in the Twitter photos. For example, the domain experts reviewed an image with artwork conveying a collection of hand drawn faces. They immediately recognized that each person drawn represented a well-known local rap artist who had been killed. Third, hand gestures in pictures were identified by domain experts as associated with specific gangs and were understood according to the message conveyed. For example, domain experts understood nuanced differences in hand gestures, including the difference between "throwing up" a gang sign (signifying affiliation or association with that gang) versus "throwing down" a gang sign (signifying disrespect towards that gang). In addition to the emergent themes, we also identified challenges with the annotation process. In some instances, domain experts admitted to unintentionally selecting the wrong code, which may reflect the time spent labeling the posts.

[1] For the analysis we exclude two tweets for which we do not have annotations from all annotators.

4.2. Findings from statistical and computational methods

As textual concepts we use the 500 most common words and emojis (computed over all 1,851 tweets), on the visual part we use a list of nine concepts (*handgun, long gun, hand gesture, joint, lean, person, tattoo, marijuana,* and *money*) which were originally defined for the purpose of training detectors for gang violence prevention and were manually annotated in all images. We run all statistical methods described in Section 3.2, using a confidence value of 0.99 for computing confidence intervals and testing significance.[1]

Table 1: Spearman correlation coefficients for psychosocial code annotations from different annotators.

annotators	aggression	loss	substance use
S 1 vs S 2	0.23	0.82	0.75
DE 1 vs DE 2	0.54	0.66	0.73
S vs DE	0.38	0.84	0.78

Annotator correlation coefficients are shown in Table 1. For *loss* and *substance use*, correlations within and between groups are all rather high (0.66 or more), indicating that for these codes, annotators label tweets in a very similar way. However, in case of *aggression* correlation coefficients are much lower. Interestingly, the lowest value of 0.23 was attained for correlation between annotations of the students.

Looking at *annotator baselines* for the different codes (Table 2) reveals that student annotators are in general far less likely to label a tweet as *aggression* as compared to domain experts (2.9% and 4.8% vs 13.4% and 20.3%). This explains how the corresponding correlation coefficient can be much lower for student annotators than for domain experts (0.23 vs 0.54), even though the disagreement baseline is lower for student annotators (5.7% vs 13.4%; see Table 3). For both other codes, baselines for all annotators are much more comparable (see last two columns of Table 2).

These findings point towards general annotator tendencies that provide important insights into the motivations for how Twitter content is labeled. For example, our domain experts may label more content as aggressive as a way to maintain safety in their community. As such, a false negative for aggression is only a minor inconvenience for the student annotators

while a false negative for the domain experts could have lethal consequences for individuals they may know. On the other hand, our student annotators may be biased towards minimizing aggression or other themes that are stereotypical or further marginalized communities of color.

Table 2: Code baselines (including confidence intervals) in percent, of student (S) and domain expert (DE) annotators for labeling tweets as the three psychosocial codes.

annotator/s	aggression	loss	substance use
S 1	2.9 (1.9-3.9)	15.9 (13.7-18.1)	11.7 (9.8-13.6)
S 2	4.8 (3.5-6.1)	15.3 (13.1-17.4)	17.2 (14.9-19.4)
S merged	6.7 (5.2-8.2)	18.0 (15.7-20.3)	12.6 (10.7-14.6)
DE 1	13.4 (11.4-15.4)	18.6 (16.3-20.9)	12.6 (10.7-14.6)
DE 2	20.3 (17.9-22.7)	11.8 (9.9-13.8)	12.3 (10.3-14.2)
DE merged	23.6 (21.0-26.1)	19.9 (17.5-22.3)	15.5 (13.3-17.6)

Table 3 and Table 4 contain *disagreement statistics* for the codes *aggression* and *substance use*. For each feature we state the total number of relevant tweets, the fraction of tweets with conflicting annotations (as difference to the respective baseline), the annotator bias and the Spearman correlation coefficient between concept presence and binary disagreement indicator. The tables only include concepts where the fraction of conflicting annotations was found to be significantly different from the respective baseline.

In the *disagreement statistics* for the code *aggression* (Table 3), for student annotators we can see that *handgun* is the most relevant concept for disagreement (with a correlation coefficient of 0.41), which intuitively makes sense. For disagreements between student annotators and domain experts, the annotator bias of 0.9 shoes that irrespective of any concept presence, in 95% of disagreement cases, domain experts voted for *aggression* while student annotators did not. The corresponding correlation coefficient of 0.40 suggests that such disagreements are often related to the presence of *hand gesture* in the image, which is in line with our findings from interviews with domain experts. Additionally, we want to point out that *hand gesture* indicates disagreement between domain experts as well, but this concept was not found to cause any conflicting annotations between student annotators. In a separate test, it did not significantly increase the likelihood of any student annotator to label a corresponding tweet as *aggression*. This means that without domain expert annotations, the relevance of *hand gesture* to *aggression* would not be visible.

Table 3: Disagreement statistics for the label aggression.

	feature	#tweets	disagr. in %	ann. bias	corr. coeff.
S 1 vs S 2	baseline	1849	5.7	+0.3	-
	(txt) 😷	69	+16.0	+0.4	0.14
	(txt) 🔫	13	+40.4	-0.4	0.15
	(img) *handgun*	164	+30.9	+0.7	0.41
	(img) *long gun*	15	+34.3	+1.0	0.13
DE 1 vs DE 2	baseline	1849	13.4	+0.5	-
	(txt) *n***az*	13	+40.4	+1.0	0.10
	(txt) *neva*	10	+56.6	+0.7	0.12
	(img) *hand gesture*	572	+15.4	+0.6	0.30
	(img) *handgun*	164	+11.6	+0.6	0.11
S vs DE	baseline	1849	19.0	+0.9	-
	(txt) *n***az*	13	+50.2	-0.8	0.11
	(txt) *neva*	10	+51.0	+1.0	0.10
	(img) *hand gesture*	572	+23.3	+0.9	0.40
	(img) *handgun*	164	+25.5	+0.9	0.20
	(+6 txt)	…	…	…	…

Table 4 lists *disagreement statistics* for *substance use*. Here, the presence of *joint* in the image of the tweet correlates with disagreement within both groups and between the two groups (coefficients 0.32, 0.27 and 0.26). For student annotators, there seems to be some additional confusion about the words “dm” and “asl” (+~50% disagreement in presence of each concept) as well as the visual presence of *lean* (+21.3% disagreement). Somewhat surprising is the finding that

handgun increases the chance of conflict between student annotators and domain experts for the label *substance use*.

Table 4: Disagreement statistics for *substance use*. All concepts with statistically significant differences to the respective baseline are included.

	feature	#tweets	disagr. in %	ann. bias	corr. coeff.
S 1 vs S 2	baseline	1849	6.6	0.8	-
	(txt) *dm*	9	+49.0	+1.0	0.14
	(txt) *asl*	7	+50.5	+1.0	0.13
	(img) *lean*	43	+21.3	+0.8	0.13
	(img) *joint*	185	+23.7	+0.9	0.32
DE 1 vs DE 2	baseline	1849	6.0	-0.1	-
	(img) *joint*	185	+19.4	+0.2	0.27
S vs DE	baseline	1849	6.2	-0.4	-
	(img) *joint*	185	+18.7	-0.9	0.26
	(img) *handgun*	164	+10.3	-0.9	0.13

Check-in's with student annotators revealed a disparate meaning-making process. For example, "dm" or direct messaging may trigger for a student annotator questions about the types of conversations that happen during a private exchange. At times the annotators misunderstood the phonetic interpretation of "asl" which in the context of our study would be used to phonetically spell a word like "as hell". The presence of a Styrofoam cup would trigger a label of an entire tweet as "lean" whereas another student annotator would not identify the entire tweet as substance use. Lastly, the socio-political interpretation of what a handgun means in an image with young African American youth informed how the student annotators labeled *substance use*.

Annotator bias. The only case where the presence of a concept significantly alters the bias for disagreement is in case of code *substance use* and visual concept *joint* for student vs domain expert disagreement. Apparently, in almost all cases (-0.9 annotator bias, i.e. around 95%) of *substance use* disagreement with a joint present in the image, student annotators voted for *substance use* and domain experts did not (as compared to the concept-independent baseline bias of around 70%). This suggests that student annotators saw *joint* as far more indicative for *substance use* than domain experts.

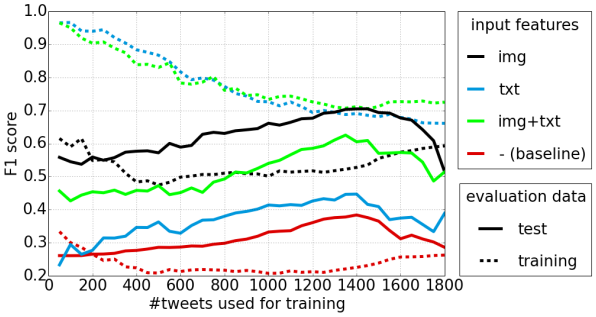

Figure 1: Performances of logistic regression models predicting disagreement between S and DE annotators for any code.

Figure 1 shows F1 scores from our experiments on *predicting disagreements* between student annotators and domain experts, comparing models that use visual, textual or both types of features. Since tweets are ordered by annotation time for this experiment, the plot visualizes the development of performances over the course of the annotation process, where at any point all current annotations are used for training and all future annotations are used as test set.

As a statistical baseline we also include performances of a system that knows the true ratio p% of items with disagreement in the evaluation data and (deterministically) classifies p% of the tweets with disagreement and p% of the tweets without disagreement as having disagreement. Note that the F1 score of this baseline is given by p/100, hence it directly describes the ratio of tweets with disagreement in the data set.

In the plot we see that, using only visual features, already after 50 tweets the prediction model achieves an F1 score of around 0.55, which is far above the respective baseline of around 0.25. For the most part, this difference remains nearly constant. The drop of performance at the end is likely due to the small number of remaining tweets for testing.

We find that for our data, adding textual concepts is detrimental to performance on unseen data, where the visual model consistently outperforms both other models and using text alone gives the worst results. Using only textual features still leads to above-baseline prediction if more than 200 tweets are used for training, but this difference remains comparatively small until the end. Considering performances on the training data clearly shows that whenever textual concepts are used as input features, prediction models apparently learn to use noise in the training set for prediction and thereby fail to generalize to the test data, a typical case of overfitting. However, this effect is getting smaller as more tweets become available for training, especially for the model that uses both visual and textual features.

Also note that the textual features we used for the experiment are more low-level and higher in magnitude as compared to our visual features. Therefore, the text modality should not be deemed generally useless for disagreement prediction based on these results.

5. Discussion

In this paper, we examine disagreements between domain experts and student annotators to consider the promise and challenge of annotating Twitter data from vulnerable and marginalized communities. Leveraging annotations from a Chicago Twitter corpus of African American and Latino youth and young adults who live in communities with high rates of gang violence, we underscore the importance of considering annotator background and involving local domain experts when analyzing Twitter data for machine learning. Specifically, nuances in culture, language, and local concepts should be considered when identifying annotators and the type of training they should receive before reviewing Twitter data. Furthermore, our findings emphasize the importance of identifying interpretation-related problems in annotation and the need for strategies on treating disagreement based on its causes.

5.1. Annotation conflicts

Much of the computer science literature focuses on eliminating disagreements between annotators, but here we argue that in the case of data from marginalized populations, some disagreement may not be negative. As we have seen, even if it is doubtful whether there really is an objective "gold standard" for the final labels, analyzing disagreements can lead to a better understanding of the domain of application. Especially in this context of more complex use-cases, if annotations are done by a few trained annotators, one can monitor their annotations and discuss disagreements as they arise, successively leading to higher quality of the annotations and a more complete overall picture.

By comparing disagreements between and within two groups of annotators, domain experts and student annotators, we uncovered critical differences in interpretation of behaviors in images on Twitter. Symbolic Interactionism theory suggests that individuals use gestures - "any part of an ongoing action that signifies the larger act or meaning" (pp. 9) to understand human behavior [5]. For example, a domain expert who lives in the same or similar community as the Twitter users under study would have a nuanced understanding of the use of the gun emoji or a specific hang gesture. They are able to situate what those specific gestures meaning within the local context, thus informing if the gesture should be determined threatening.

When gestures are interpreted incorrectly, we risk inflicting a detrimental and compounded impact on the current and future experiences for marginalized users already experiencing the results of systematic oppression. Patton et al. [16] uncovered distinct differences in how police use social media as evidence in criminal proceeding. For example, the misinterpretation of gestures made by young African American men on Facebook led to the arrest of over 100 young Black men in New York City, some of whom were not involved with the crime under question [22]. Conversely, social media threats made by a White male, Dylann Roof, who killed nine African American church-goers in Charleston, South Carolina, went undetected by law enforcement. In addition, Safiya Noble [11] warns us that biases unchecked in the labeling of images on google reinforce racist stereotypes.

Understanding and analyzing disagreements benefitted our annotators. At the micro level, this process pushed our student annotators to redefine labels that could lead toward providing a user with additional supports and resources. At the macro level our processes forced us to consider how applying the wrong label could further criminalize an already stigmatized population. For example, interpreting a *hand gesture* that represents gang association in case of *aggression* only became evident after consulting with experts, so the "true" meaning of *hand gesture* would have been missed by our student annotators. This implies that the common strategy of adding more non-expert annotators would likely not have revealed this aspect either.

Luckily, we found that computational models can learn to predict disagreement between social work annotators and domain experts from rather few samples when using suitable features for the prediction. In practice this can potentially be useful for better leveraging community members' expertise by automatically selecting critical examples for expert annotation. Essentially, this would mean adopting an active learning paradigm for selectively collecting annotations, similar to [24], but instead of focusing on detectors, expert annotations would be selected in order to train annotators or content moderators.

5.2. Role of domain experts in annotation

Domain expertise is vital to annotating Twitter data from marginalized communities. In the study of gang-related concepts on Twitter, we hired domain experts to perform several functions. First, we leveraged insights from domain experts to train student annotators on

nuances in language, culture and context that are embedded in the text and images in the Twitter posts. Second, domain experts separately annotated Twitter posts from users in their own community, which allowed us to compare their annotations to graduate student annotations. These annotations help us understand how people from the community naturally view Twitter posts using their experience and expertise. Third, we interviewed them to understand how they made decisions and what informed the labels they assigned to images. Interviews with the domain experts revealed critical concepts like handmade cards or recognizing people which were visible in the images, but not captured by our visual concepts. The critical concepts are not frequent and thus challenging to detect using statistical or automatic methods. Even if it were possible to detect these concepts it would be impossible to find out the extent to which a hand gesture is important without interviewing the domain experts.

Domain experts and student annotators engage the annotation process differently. Our domain experts have more intuitive and instinctive interpretations of Twitter posts because those posts reflect their everyday lived experiences and how they interpret their community. Conversely, the student annotators are trained to annotate using a detailed process, specifically considering alternative meanings and interpretations because they do not have the same contextual experiences. Weighing the differences between domain experts and student annotators should be informed by the research question and specific tasks. In this study, domain experts provide a nuanced understanding of language and behavior (e.g. hand gestures) that our student annotators would only understand if they had the same lived experiences. Our student annotators pushed us to consider the broader ethical challenges that come with annotating Twitter data from African American and Latino youth and young adults.

5.3. Ethical considerations

As researchers who study gang-related content on Twitter, we understand our ethical obligations to ensure that our work does not further criminalize or stigmatize African American users in our sample. To protect the users in our corpus, we will only publish specific parts of the statistical features to prevent the ability to trace our users. Given the popular use of social media to monitor communities for potential acts of violence, this study underscores the importance of domain expertise and studying disagreement to highlight challenges in perception and interpretation of Twitter data from marginalized communities.

6. Future Directions

This work has implications for the development of and training for content moderation at social media platforms. Companies like Facebook and Twitter might consider training sessions where disagreements between moderators are identified and reviewed to identify moderator bias and gain additional contextual and cultural insights that may inform how they make decisions about removing content.

As another step, we plan to apply our methods for annotator perspective analysis in several other scenarios. First, we plan to use annotations from different datasets, such as text-only tweets of gang-involved youth [4] or even annotations of image captions on Flickr collected over crowdsourcing platforms [3]. Second, we want to test how generalizable these methods are by using them to evaluate misclassifications of machine learning algorithms, which can be seen as disagreement between a detector and human annotators, or to compare functioning of multiple automatic methods.

References

[1] Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163-222). Springer, Boston, MA.

[2] Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social research update*, *33*(1), 1-4.

[3] Blandfort, P., Karayil, T., Borth, D., & Dengel, A. (2017, October). Image Captioning in the Wild: How People Caption Images on Flickr. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes* (pp. 21-29). ACM.

[4] Blevins, T., Kwiatkowski, R., Macbeth, J., McKeown, K., Patton, D., & Rambow, O. (2016). Automatically processing tweets from gang-involved youth: towards detecting loss and aggression. In Proceedings of *COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2196-2206).

[5] Blummer, H. (1969). Symbolic Interactionism. *Perspective and Method*. University of California Press. Berkeley, CA.

[6] Brock, A. (2018). Critical technocultural discourse analysis. *New Media & Society*, 20(3), 1012-1030.

[7] Broussard, M. When Cops Check Facebook. *The Atlantic*, 2018. https://www.theatlantic.com/politics/archive/2015/04/when-cops-check-facebook/390882/.

[8] Frey, W. R., Patton, D. U., Gaskell, M. B., & McGregor, K. A. (2018). Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data. *Social Science Computer Review*, 0894439318788314.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[10] Miltsakaki, E., Joshi, A., Prasad, R., & Webber, B. (2004). Annotating discourse connectives and their arguments. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*.

[11] Noble, Safiya. (2018) Algorithms of Oppression. How Search Engines Reinforce Racism. *New York University Press*. New York.

[12] Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*(1), 71-106.

[13] Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. *Language Resources and Evaluation*.

[14] Passonneau, R. J. (2004, May). Computing Reliability for Coreference Annotation. In *LREC*.

[15] Patton, D.U., Blandfort, P., Frey, W.R., Schifanella, R., & McGregor, K. (under review). VATAS: An open-source web platform for visual and textual analysis of social media.

[16] Patton, D. U., Brunton, D. W., Dixon, A., Miller, R. J., Leonard, P., & Hackman, R. (2017). Stop and Frisk Online: Theorizing Everyday Racism in Digital Policing in the Use of Social Media for Identification of Criminal Conduct and Associations. *Social Media + Society*, *3*(3), 2056305117733344.

[17] Patton, D.U., Frey, W.R., McGregor, K.A., Lee, F.T., McKeown, K.R. (under review). Contextual analysis of social media: a qualitative approach to eliciting context in social media posts with natural language processing.

[18] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, *11*(Apr), 1297-1322.

[19] Reidsma, D., & Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, *34*(3), 319-326.

[20] Roberts, S.T. (2016). Commercial content moderation: Digital laborers' dirty work. In Noble, S.U. and Tynes, B. (Eds.), *The intersectional internet: Race, sex, class and culture online* (2016), 147-159. Download here: https://illusionofvolition.com/publications-and-media/

[21] Rodrigues, F., Pereira, F., & Ribeiro, B. (2013). Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, *34*(12), 1428-1436.

[22] Speri, Alison (2014). The Kids Arrested in the Largest Gang Bust in New York City History Got Caught because of Facebook. *VICE*. Retrieved here: https://news.vice.com/

[23] University of Chicago Crime Lab (2016). Gun Violence in 2016. Chicago, IL. Retrieved from: http://urbanlabs.uchicago.edu/attachments/store/2435a5d4658e2ca19f4f225b810ce0dbdb9231cbdb8d702e784087469ee3/UChicagoCrimeLab+Gun+Violence+in+Chicago+2016.pdf

[24] Yan, Y., Rosales, R., Fung, G., & Dy, J. G. (2011, June). Active learning from crowds. In *ICML* (Vol. 11, pp. 1161-1168).

[25] Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., ... & Dy, J. (2010, March). Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 932-939).