Evaluating Annotation Consistency in Offensive Language Detection: A Data Analytics Approach on the TweetEval Dataset

Dr. Fabeela Ali Rawther
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
fabeelaalirawther@amaljyothi.ac.in

Abhinay A K
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
abhinayak2025@cs.ajce.in

Anagha Tess B
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
anaghatessb2025@cs.ajce.in

Alan Joseph
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
alanjoseph2025@cs.ajce.in

Adham Saheer
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
adhamsaheer2025@cs.ajce.in

***Abstract*— Most machine learning models are not only highly dependent on difficult datasets but also on the quality of labeled data they are trained on, especially for offensive content detection. In this paper, we study the TweetEval dataset to provide a comparison of its ground truth with manually annotated labels; inter-annotator agreements are applied here as a metric for assessing the consistency of annotation. Cohen's Kappa coefficient is used to quantify how much each pair of annotators agreed and where they differed. In-depth examination of missed classifications demonstrates other difficulties with manual labelling: subjective interpretation, context dependency, and annotator bias. The insights gathered demonstrate how manual annotation can have positive and negative effects on further model training practices, highlighting the importance of standardized annotation guidelines. In their actions, the findings contribute to enhancing offensive content detection models by advocating dataset reliability and the reduction of inconsistencies in labeling.**

***Index Terms*—TweetEval Dataset,Annotation Consistency, Inter-Annotator Agreement,Cohen's Kappa, Fleiss' Kappa,Dataset Reliability, Text Classification,Natural Language Processing (NLP), Offensive Language Detection, Hybrid Models,Annotator Bias**

I. INTRODUCTION

A. Background and Motivation:

The rise of social media has raised concerns globally of the proliferation of not only offensive content but also harmful content on the Internet. Such concerns are becoming a target, relying on annotated datasets, for machine-imparted detection and classification of offensive language[1]. The performance of these models is highly contingent upon how accurate and consistent their annotations were within the dataset. While datasets pose their own set of predefined labels, there exists the miscellaneous nature of how the same piece of content is interpreted by human annotators: this can vary and lead to inconsistency. Inconsistency might make such models perform differently and bring in a bias; therefore, it's very important to consider such a reliability check.

B. Challenges in Manual Annotation:

One of the main issues with offensive content detection is the human subjectivity associated with annotation. Different annotators could have different opinions on what constitutes offensive language based on personal experiences, cultural background and language-related nuances[3]. This would create a disparity between the truthful labels and the labels written by humans, discrediting the reliability of the exact dataset on which the analysis was done. The TweetEval dataset, an applicable and reliable benchmark for tweet classification tasks, shows similar inconsistencies, thus being a great case by which the quality of annotation can be evaluated[4].

C. Objectives:

This research is concerned with examining alignment between actual labels and human annotations within the TweetEval dataset, the main goal being to assess annotation consistency through inter-annotator agreement metrics. Measures like Cohen's Kappa and Fleiss' Kappa are used to measure agreement levels among the annotators. Through the determination of annotation inconsistencies and possible biases, this work seeks to advance dataset curation processes and the reliability of AI-based content moderation systems. The major contributions of this research are:

- Refining annotation guidelines for enhancing dataset consistency and reliability.
- Improving the fairness and resilience of AI-based moderation tools through resolution of subjectivity in annotations.
- Giving a structure to quantifying annotation consistency, assisting with future studies on dataset curation and assessment.

II. Related Works

The reliability of labeled datasets is a fundamental concern in natural language processing (NLP), particularly in tasks involving subjective interpretations such as offensive content detection. Previous studies have extensively explored inter-annotator agreement as a measure of dataset quality, emphasizing the need for consistency in human annotations to ensure robust model performance.

Artstein and Poesio [1] provided a comprehensive review of inter-coder agreement in computational linguistics, highlighting statistical measures such as Cohen's Kappa and Fleiss' Kappa for assessing annotation reliability. Their work underscored the challenges posed by subjective interpretation in linguistic tasks and the necessity of well-defined annotation guidelines. Bayerl and Paul [2] further examined the factors influencing inter-coder agreement, concluding that annotation consistency is highly dependent on task complexity, annotator expertise, and the clarity of annotation instructions.

Carletta [3] explored the use of the Kappa statistic in classification tasks, demonstrating its effectiveness in evaluating agreement beyond chance. Meanwhile, Byrt et al. [4] discussed the impact of bias and prevalence on Kappa values, noting that variations in annotator perceptions can significantly affect agreement scores. These studies collectively highlight the limitations of inter-annotator agreement metrics and emphasize the need for supplementary methods to account for annotation subjectivity.

In the context of dialogue systems and conversational AI, Artstein et al. [5,6] investigated the feasibility of structured annotation schemes in tactical questioning dialogues. Their findings suggest that even with predefined annotation frameworks, achieving high agreement remains challenging due to the inherent variability in human interpretation. Similarly, Bennett et al. [7] examined the role of limited questioning in communication and its effect on classification reliability, which is particularly relevant to offensive content detection where context plays a crucial role in annotation decisions.

III. Datasets and Analysis

A. Dataset Annotation

The dataset used in this study is one of the benchmark datasets called TweetEval intended for different tweet classification tasks such as offensive content detection. It contains tweets that are pre-labeled in specific categories in order to train and test models. Further a manual annotation was done on it using a simple binary classification, which is either offensive content (1) or non-offensive content (0). Fig.1 shows the dataset with true label and predicted label.

	Test Statement	True Label	Predicted Label
2	@USER too much thoughts inside his headdd we can't ever	0.305953979	1
3	First time I heard his name in camp, he seems to be the forg	0.1942929626	1
4	When I go to drink with Tsubaki he would always fall asleep	0.2953295936	0
5	@USER His ass need to stay up ð□□□ð□□□	0.8333493191	1
6	most important tweet of the day : Fuck Donald Trump and hi	0.5645274354	1
7	You wanna leave? then feel free cus I promise I wonâ□□t st	0.6649215868	1
8	@USER lmfao itâ□□s gotta be a highlight video of him gettir	0.3436436293	0
9	@USER -10/10 Will sooner put him on his ass in a duel thai	0.7447527517	0
10	@USER Paul Hollywood - I wouldnâ□□t hear a word he sai	0.3016097604	0

Fig. 1. Dataset

The manually annotated data further was compared with the true reference provided in the TweetEval benchmark. Areas in which discrepancies were observed between annotations were assessed and examined for aspects of mis-classification. Common challenges faced in annotation included language interpretation variance, sarcasm, and implicit offensive content. These complexities highlighted the intricacy of moderation tasks and the need for strictly defined set guidelines for annotation. This study also sheds further light on the reliability of annotation and its quality within a data set by providing evidence of the differences between manual annotations and existing data set labels. The outcome of these results would further augment the dexterity of annotation procedures and the coherence of information used in AI methods for content moderation.

B. Inter-Annotator Agreement Analysis

Inter-annotator agreement is a measurement of the various kinds of consistencies across manual annotations and, thus, the extent to which annotators agree on labeling decisions. In this study, agreement is quantified by means of statistical metrics such as Cohen's Kappa and Fleiss' Kappa dealing with the level of agreement.

Cohen's Kappa (): Measures pairwise agreement between two annotators while taking into account the chance of agreement by pure random guess. A value of above 0.75 is usually associated with strong agreement, while below 0.40 would indicate a poor agreement.

Fleiss' Kappa (F): Extends the Cohen Kappa to a scenario involving more than two annotators and yields one total measure of consistency. This metric works especially well for multi-label annotation tasks, like those in the TweetEval dataset.

IV. Discussion and Findings

Annotation inconsistencies can affect model performance, that is, ambiguity and misclassification among instances, leading to reduced accuracy during binary classifications. Such differences in agreement give a sense of how inconsistent the labels are, resulting in incorrect learning. Furthermore, agreement variations might also indicate annotator bias, since labeling builds upon culture and personal beliefs, which may introduce systematic errors. Something that is thought of as offensive by one annotator may look neutral to another, affecting

the prediction model. Clear guidelines and multiple rounds of reviews will be needed to tackle such biases, to raise label consistency and guarantee a more reliable model.

The evaluation of inter-annotator agreement indicates a high degree of consistency in manual annotations, as seen by the computed agreement scores. A average Cohen's Kappa of 0.80 gives good evidence for the strong agreement of the annotators and strengthens the reliability of the labeled dataset. The small differences in agreement also demonstrate the partially subjective nature of offensive content classification. The lowest observed agreement, in general, was 0.77, whereby the annotators disagreed due to contextual ambiguity.Fleiss' Kappa was then computed to show agreement consistency; it produced a value of 0.79, which correlates to high agreement between the different annotators involved.

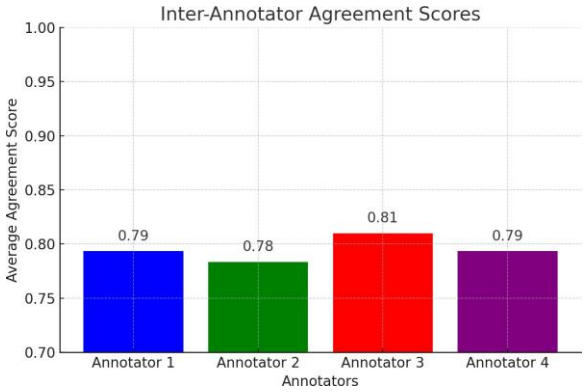

Fig. 2. Inter Annotator Agreement Scores

The Fig.2 bar chart shows the inter-annotator agreement scores for four annotators, dubbed by their average Cohen's Kappa values. These scores are an indication of how consistent or effective their annotations on offensive content classification are with those of others that contributed to it as well.In general, key observations include the following: agreement scores by annotators fall between 0.78 and 0.81, showing an important agreement among those annotators of the dataset, this suggests that they understood similar meanings of the term offensive content, as far as highest agreement is concerned, Annotator 3 scored 0.81, this shows that Annotator 3 interpretation was probably aligned with other annotators when labeling,hence Annotator 3 seems the most consistent.The lowest agreement score, that is Annotator 2, experienced the least agreement might imply slightly different interpretations on what constitutes offensive content, probably due to subjective biases or understanding of context.

Differences in Annotator decisions open questions on annotation guidelines calls for improvement and maybe more posttraining to make agreement better.Notwithstanding the discrepancies, high agreement suggests much greater conformity in best evaluation practices for offline annotation in the lengthy build-up since general formal acceptance of annotator input contained here. Theory and testing explored here then corroborates results of the TweetEval dataset as reflectively high in annotation reliability.

V. Conclusion and Future Works

This work sheds light on the importance of annotation consistency to enhance the effectiveness of their offensive content detection model developed on the TweetEval dataset. The inter-annotator agreement analysis showed a substantial agreement between the annotators given Cohen's Kappa and Fleiss' Kappa. Nevertheless, their differentials also allude to several underlying hurdles: personal interpretations, cultural influences, and the inbuilt subjectivity in the classification of offensive language. More importantly, these inconsistencies lead to certain unavoidable biases in the dataset, resulting in potential misclassification by the classification model and its poor performance. The annotation also becomes complex as contextual ambiguity alters the meaning of the utterance as per external factors, such as tone, intent, or prior discourse.

This study spots annotator bias as one of the major challenges, resulting from individual comments on what constitutes offensive language. Such biases can weaken the quality of annotations and affect the fairness and generalizability of the resulting models in the absence of a standardized guideline and sufficient training. Based on these principles, it is in the interest of various stakeholders to tackle these issues in order to improve dataset quality and make sure that AI-driven content moderation systems function better and are more equitable.

In the future, research will look toward improving the annotation protocols through clear labeling criteria as well as training the annotators further to reduce subjectivity. Automated annotation techniques that include weak supervision, active learning, and crowdsourcing strategies are to enhance dataset reliability with less manual effort. Also, there is an opportunity of extending the analysis to other offensive language datasets through cross-dataset comparisons for further understanding of the consistency of annotations in differing contexts. Improvements made in this line will contribute toward the development of trustworthy and relatively unbiased datasets toward the detection of offensive content in natural language processing applications.

References

[1] R. Artstein, M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[2] P. S. Bayerl, K. I. Paul, "What determines inter-coder agreement in manual annotations? A meta-analytic investigation," *Computational Linguistics*, vol. 37, no. 4, pp. 699–725, 2011.

[3] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[4] T. Byrt, J. Bishop, J. B. Carlin, "Bias, prevalence and kappa," *Journal of Clinical Epidemiology*, vol. 46, no. 5, pp. 423–429, 1993.

[5] R. Artstein, S. Gandhe, J. Gerten, A. Leuski, D. Traum, "Semi-formal evaluation of conversational characters," in *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, O. Grumberg, M. Kaminski, S. Katz, S. Wintner (eds.), Lecture Notes in Computer Science, vol. 5533, Springer, Heidelberg, pp. 22–35, 2009.

[6] R. Artstein, M. Rushforth, S. Gandhe, D. Traum, A. Donigian, ”Limits of simple dialogue acts for tactical questioning dialogues,” in *Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Barcelona, Spain, pp. 1–8, 2011.

[7] E. M. Bennett, R. Alpert, A. C. Goldstein, ”Communications through limited questioning,” *Public Opinion Quarterly*, vol. 18, no. 3, pp. 303–308, 1954.