IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 April 2024, accepted 14 May 2024, date of publication 20 May 2024, date of current version 29 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402809

RESEARCH ARTICLE

ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks

ARBI HAZA NASUTION[1] **AND AYTUĞ ONAN**[2]
[1]Department of Informatics Engineering, Universitas Islam Riau, Pekanbaru, Riau 28284, Indonesia
[2]Department of Computer Engineering, College of Engineering and Architecture, İzmir Kâtip Çelebi University, 35620 İzmir, Turkey

Corresponding author: Aytuğ Onan (aytug.onan@ikcu.edu.tr)

This work was supported in part by Universitas Islam Riau.

ABSTRACT This research paper presents a comprehensive comparative study assessing the quality of annotations in Turkish, Indonesian, and Minangkabau Natural Language Processing (NLP) tasks, with a specific focus on the contrast between annotations generated by human annotators and those produced by Large Language Models (LLMs). In the context of NLP, high-quality annotations play a pivotal role in training and evaluating machine-learning models. The study encompasses three core NLP tasks: topic classification, tweet sentiment analysis, and emotion classification, each reflecting a distinct aspect of text analysis. The research methodology incorporates a meticulously curated dataset sourced from a variety of text data, spanning diverse topics and emotions. Human annotators, proficient in the Turkish, Indonesian, and Minangkabau language, were tasked with producing high-quality annotations, adhering to comprehensive annotation guidelines. Additionally, fine-tuned Turkish LLMs were employed to generate annotations for the same tasks. The evaluation process employed precision, recall, and F1-score metrics, tailored to each specific NLP task. The findings of this study underscore the nuanced nature of annotation quality. While LLM-generated annotations demonstrated competitive quality, particularly in sentiment analysis, human-generated annotations consistently outperformed LLM-generated ones in more intricate NLP tasks. The observed differences highlight LLM limitations in understanding context and addressing ambiguity. This research contributes to the ongoing discourse on annotation sources in Turkish, Indonesian, and Minangkabau NLP, emphasizing the importance of judicious selection between human and LLM-generated annotations. It also underscores the necessity for continued advancements in LLM capabilities, as they continue to reshape the landscape of data annotation in NLP and machine learning.

INDEX TERMS Annotation quality, emotion classification, Indonesian language processing, language models, low-resource languages, natural language processing, sentiment analysis, topic classification, Turkish language processing.

I. INTRODUCTION

The field of Natural Language Processing (NLP) and machine learning stands at the intersection of human language and artificial intelligence, with profound implications for how we interact with technology. In this realm, data annotations play a pivotal role, serving as the cornerstone for building and evaluating NLP models [1], [2]. These annotations, encompassing everything from sentiment labels to named entities, imbue unstructured text data with meaning and structure, making it comprehensible to machines. The quality and reliability of these annotations hold the key to the

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Kavak.

performance and utility of NLP applications, from chatbots and virtual assistants to machine translation systems and information retrieval engines [3].

Many NLP applications hinge on the availability of high-quality labeled data, often requiring vast amounts of annotated text to train, fine-tune, or evaluate machine learning models [4]. In supervised tasks like sentiment analysis or text classification, annotated data serves as the compass guiding models to distinguish between positive and negative sentiment, categorize news articles, or identify spam emails [5]. In unsupervised tasks, such as topic modeling and document clustering, annotated data provides the benchmarks for evaluating the coherence and relevance of automatically generated clusters and topics [6].

The importance of data annotations in NLP and machine learning extends beyond practical applications into the very heart of the field's research and development. Annotations act as the ground truth against which models are tested, refined, and advanced. They enable researchers to push the boundaries of language understanding, to train models that can converse like humans, and to uncover the intricate structures hidden within texts [7]. However, achieving high-quality annotations is not without its challenges. It necessitates a careful balance between linguistic expertise, domain knowledge, and meticulous annotation guidelines [8]. Moreover, the traditional approach to annotation, reliant on human annotators, often poses constraints related to cost, scale, and consistency [9], [10].

In recent years, a transformative shift has emerged with the ascendancy of Language Models (LLMs) like GPT-3 and its counterparts [11]. These sophisticated models, driven by neural architectures and fueled by the breadth of extensive pre-training, represent the apex of language understanding [12]. They have redefined the very essence of NLP, serving as the bedrock upon which modern language applications are built. LLMs empower machines to decipher the intricate nuances of human language, navigating the labyrinth of syntax, semantics, and context with unparalleled precision [13]. From sentiment analysis to machine translation, LLMs are the catalysts propelling NLP to new heights, ushering in an era of more accurate, context-aware, and versatile language applications [14]. These models, trained on extensive corpora of text data, possess the remarkable ability to generate coherent text and, notably, to produce annotations for various NLP tasks [15]. This paradigm shift in annotation generation brings forth a new era of scalability and efficiency, with the potential to alleviate some of the challenges associated with human annotation [16].

The growing role of LLMs in data annotation demands a thorough examination of their capabilities and limitations. While they hold the promise of speed and scalability, questions about their ability to generate high-quality annotations persist. Can LLMs consistently produce annotations that match or surpass the quality of human-generated ones? Do their annotations exhibit variations depending on the nature and complexity of the NLP task at hand?

The primary motivation of this study is two-fold. Firstly, we aim to contribute to the ongoing discourse on the reliability and applicability of LLM-generated annotations in Turkish, Indonesian, and Minangkabau NLP tasks. While LLMs offer undeniable advantages, their limitations in context comprehension, ambiguity resolution, and domain-specific nuances require rigorous scrutiny. Secondly, we endeavor to address the practical implications of annotation source selection in the era of LLMs. As the demand for annotated data escalates, understanding when and where LLM-generated annotations can be confidently employed versus when human expertise remains irreplaceable becomes paramount.

This study systematically explores LLM-generated annotations in Turkish, Indonesian, and Minangkabau NLP tasks. In Section II, prior research on data annotation and LLMs have been briefly reviewed. In Section III, we outline our approach, including data selection and annotation process. We also detail the dataset partitioning and LLM configuration in Section III. In Section IV, we present results and comparisons with metrics and examples. In addition, we interpret findings, focusing on annotation quality factors. In Section V, we summarize key findings, emphasizing significance and future directions.

II. RELATED WORK

The rapidly evolving domain of text annotation and classification has seen large language models (LLMs) like ChatGPT emerge as valuable tools. Their ability to execute zero-shot tasks and provide meaningful annotations has become a focal point of contemporary research. This section collates pertinent research endeavors that probe into the intricacies of such applications.

Kuzman et al. [17] conducted a study to investigate the potential of ChatGPT for zero-shot text classification, specifically focusing on automatic genre identification. Their research compared ChatGPT's performance with a fine-tuned multilingual language model on English and Slovenian datasets. The results demonstrated ChatGPT's effectiveness in genre identification, raising questions about the need for laborious manual annotation campaigns, even for smaller languages.

Laskar et al. [18] proposed a methodology for cleaning the Debatepedia dataset using ChatGPT to improve query-focused abstractive summarization. They found that ChatGPT's annotations enhanced query relevance and summary quality, offering a valuable resource for research.

Ollion et al. [19] conducted a systematic literature review to assess the merits and limitations of zero-shot text annotation using models like ChatGPT. They analyzed various articles in the human and social sciences and found mixed results in terms of performance. They emphasized the need for further exploration while acknowledging important questions related to reproducibility, privacy, and language diversity.

Gilardi et al. [20] investigated the use of ChatGPT for text annotation tasks and demonstrated its superiority

over crowd-workers in several annotation tasks, including relevance, stance, topics, and frames detection. They highlighted the cost-effectiveness of ChatGPT and its potential to improve text classification efficiency.

Ostyakova et al. [21] explored the viability of ChatGPT in generating data for complex linguistic annotation tasks, specifically focusing on speech functions in open-domain conversations. They compared data generated by ChatGPT with manually annotated datasets, shedding light on the use of large language models in linguistic annotation.

In another study, Ostyakova et al. [22] proposed a semi-automated method for annotating open-domain conversations with speech functions using hierarchical guidelines and ChatGPT. Their study compared annotation results from experts, crowd-workers, and ChatGPT, showcasing the potential of large language models in complex discourse annotation.

Koptyra et al. [23] investigated the possibility of using ChatGPT to automatically annotate texts with emotions and presented various datasets, including CLARIN-Emo and ChatGPT-Emo. They discussed the advantages and limitations of manual annotation compared to ChatGPT-generated data.

Vujinović et al. [24] explored ChatGPT's effectiveness in annotating datasets for training instructor models in intelligent tutoring systems. They introduced a novel dataset annotation methodology and demonstrated ChatGPT's potential as an alternative to human experts, addressing the challenges of dataset creation in the pedagogical context.

Belal et al. [25] examined ChatGPT's use as a tool for data labeling in sentiment analysis tasks. They reported significant improvements in accuracy compared to lexicon-based methods, showcasing ChatGPT's potential in annotating sentiment-related datasets.

Reiss [26] conducted a study to assess the consistency and reliability of ChatGPT's zero-shot capabilities for text annotation and classification. Their findings highlighted variations in ChatGPT's output based on model parameters and prompts, emphasizing the importance of thorough validation.

Törnberg [27] evaluated ChatGPT-4's accuracy, reliability, and bias in annotating political Twitter messages, comparing it to expert classifiers and crowd-workers. The study revealed ChatGPT-4's potential to outperform human classifiers, particularly in tasks requiring contextual knowledge and inferences.

Huang et al. [7] explored ChatGPT's utility in providing natural language explanations for the detection of implicit hate speech. They conducted user studies to evaluate the quality of ChatGPT-generated explanations compared to human-written explanations, highlighting the potential and limitations of ChatGPT in this context.

Alizadeh et al. [28] examined the performance of open-source Large Language Models (LLMs) in text annotation tasks and compared them to proprietary models like ChatGPT and human-based services. Their findings indicated the cost-effectiveness and competitive potential of open-source LLMs in specific tasks.

Zhu et al. [29] investigated ChatGPT's potential to reproduce human-generated label annotations in social computing tasks. They examined ChatGPT's performance in relabeling seminal datasets and discussed the challenges and opportunities in utilizing ChatGPT for data annotation.

Wang et al. [30] explored the use of GPT-3 as a low-cost data labeler for training other models across various NLP tasks. They demonstrated that GPT-3's labels could achieve comparable performance to human-generated labels, presenting a cost-effective data labeling methodology.

While the aforementioned studies have made significant strides in exploring the capabilities of large language models (LLMs) like ChatGPT in text annotation and classification, several notable gaps remain in the literature. These gaps form the motivation for our current study: Diverse NLP Tasks in Turkish, Indonesian, and Minangkabau: While there is substantial research on English NLP tasks with LLMs, there is a dearth of studies focusing on Turkish, Indonesian, and Minangkabau NLP tasks. Our study fills this gap by specifically evaluating LLM-generated annotations in Turkish, Indonesian, and Minangkabau, expanding the scope of research to languages beyond English.

Comprehensive Annotation Quality Assessment: Prior research has often emphasized the effectiveness of LLMs in generating annotations but has not comprehensively assessed the quality of these annotations. Our study extends this research by rigorously comparing the quality of LLM-generated annotations with human-generated annotations across multiple NLP tasks, providing a more nuanced evaluation.

Comparison with Human Experts: While some studies have compared LLMs with crowd-workers, there is a paucity of research that assesses LLMs in comparison to human experts. Our study bridges this gap by evaluating the performance of LLM-generated annotations against expert annotations, shedding light on the potential role of LLMs as alternatives to human experts.

By addressing these gaps in the existing literature, our study seeks to contribute a comprehensive assessment of the quality, and reliability surrounding LLM-generated annotations in the context of Turkish, Indonesian, and Minangkabau NLP tasks, thus advancing the understanding of the capabilities and limitations of LLMs in text annotation and classification.

In summary, while existing literature provides valuable insights into NLP applications across various languages and contexts, there remains a significant gap in research focusing on low-resource languages like Turkish, Indonesian, and Minangkabau. Our study addresses this gap by not only exploring the efficacy of Large Language Models (LLMs) in these languages but also by providing a comparative analysis with human-generated annotations. This approach is particularly innovative as it sheds light on the nuanced capabilities and limitations of both human and machine

annotations in less commonly studied linguistic contexts. This focus on low-resource languages is crucial for the advancement of equitable and inclusive NLP research, ensuring that the benefits of technological advancements are accessible across diverse linguistic landscapes.

III. MATERIALS AND METHODS

This section delineates the systematic approach underpinning our research. We commence with an exploration of the dataset, outlining its origins, scale, and textual composition. This is followed by an in-depth discussion on the three core Turkish NLP tasks we focused on: topic classification, tweet sentiment analysis, and emotion classification and two core Indonesian NLP tasks we focused on: tweet sentiment analysis and emotion classification. To ensure consistency and reliability, we provide a snapshot of the guidelines crafted for human annotators. Lastly, we present specifics regarding the fine-tuned Turkish Large Language Models (LLMs), detailing their architecture and training nuances.

A. DATASETS

Our research employed three distinct Turkish datasets, each tailored for a specific NLP task-topic classification, tweet sentiment analysis, and emotion classification, two distinct Indonesian datasets, and two distinct Minangkabau datasets each tailored for a specific NLP task-tweet sentiment analysis and emotion classification as shown in Table 1.

1) DATASET FOR TOPIC CLASSIFICATION (DTC)

Assembled from various proprietary in-house collections and enriched with articles from leading Turkish newspapers (Cumhuriyet, Hürriyet, Sabah), the DTC offers 60,000 samples, representing a comprehensive scope of contemporary Turkish texts. The content composition includes 50% news articles (covering current events, international affairs, economics, and societal trends), 30% from reputed local academic journals (spanning humanities to natural sciences), and 20% excerpts from modern Turkish literature. For this dataset, the class labels include Current Affairs, Geopolitics, Economics, Societal Issues, Humanities, Social Sciences, Natural Sciences, Prose, and Poetry.

2) DATASET FOR TWEET SENTIMENT ANALYSIS (DTSA)

Exclusively derived from the Twitter API, this dataset homes 50,000 tweets, presenting a snapshot of prevailing public sentiments. Emphasis during collection was placed on original Turkish content, with retweets and non-Turkish entries filtered out. The dataset prioritized tweets with clear sentiment indicators while ambiguous or neutral tweets were excluded for analytical clarity. For this dataset, the class labels include Positive, Negative, and Neutral.

3) DATASET FOR EMOTION CLASSIFICATION (DEC)

Anchored in contemporary Turkish literature, the DEC consists of 40,000 excerpts from a variety of genres penned by prominent Turkish writers over the last two decades. Selection criteria prioritized emotional expressiveness and consistency in content length (ranging between 50 to 200 words). The chosen samples exhibit clear emotional undertones, tying them to pre-established emotional categories, thus streamlining subsequent annotation efforts. For this dataset, the class labels include Happiness, Sadness, Anger, Fear, Surprise, Disgust, and Neutral. The basic descriptive information regarding the three datasets has been summarized in Table 1.

4) INDONESIAN DATASET FOR TWEET SENTIMENT ANALYSIS (IDTSA)

This sentence-level sentiment analysis dataset published by IndoNLU [31] is a collection of comments and reviews in Indonesian obtained from multiple online platforms. The text was crawled and then annotated by several Indonesian linguists to construct the dataset. There are three possible sentiments on the dataset: Positive, Negative, and Neutral.

5) INDONESIAN DATASET FOR EMOTION CLASSIFICATION (IDEC)

Collected from the social media platform Twitter using Twitter Streaming API for about 2 weeks, starting from June 1, 2018 until June 14, 2018 [32], a publicly available Indonesian emotion classification dataset consists of 4,403 Indonesian colloquial language tweets, covering five different emotion labels include Love, Happiness, Sadness, Anger, and Fear. The dataset define emotion in personal tweets, thus, the tweets from news portal, government office, and commercial promotion are excluded.

6) MINANGKABAU DATASET FOR TWEET SENTIMENT ANALYSIS (MDTSA) AND MINANGKABAU DATASET FOR EMOTION CLASSIFICATION (MDEC)

Using a bilingual Indonesian-Minangkabau dictionary with a total of 4,680 translation pairs, IDTSA is translated into MDTSA and IDEC is translated into MDEC.

Our Dataset for Topic Classification (DTC) was meticulously sourced from authorized repositories and digital libraries, ensuring ethically procured articles and journals. Following the standard preprocessing techniques, we removed headers, footers, special characters, and applied tokenization to ensure a uniform dataset. The dataset boasts 30,000 samples from news articles, with 12,000 (40%) dedicated to current affairs, 9,000 (30%) to geopolitics, 6,000 (20%) to economics, and the remaining 3,000 (10%) addressing societal issues. Alongside, the dataset incorporates 18,000 samples from academic journals, distributed equally among humanities, social sciences, and natural sciences, each constituting 6,000 samples or 33.3%. Furthermore, the DTC includes 12,000 literature excerpts with 7,200 (60%) from prose and 4,800 (40%) from poetry.

For the Dataset for Tweet Sentiment Analysis (DTSA), tweets were responsibly acquired via Twitter's API, adhering strictly to its user privacy guidelines. The preprocessing

TABLE 1. The basic descriptive information for datasets.

Dataset	DTC	DTSA	DEC	IDTSA	IDEC	MDTSA	MDEC
Total Samples	60,000	50,000	40,000	12,760	4,403	12,760	4,403
Average Word Count	512	18	125	32	29	32	29
Standard Deviation (Word Count)	50	4	30	21	9	21	9
Median Word Count	495	18	123	28	28	28	28
Max Word Count	615	28	212	110	58	110	58
Min Word Count	423	10	52	1	2	1	2

included the removal of URLs, mentions, hashtags, and emojis, and incorporated stemming and stopword removal for refinement. The sentiment distribution yielded 18,000 positive tweets (36%), 20,000 neutral (40%), and 12,000 negative (24%). Challenges arose due to tweets' succinct nature, necessitating specialized tokenization, and the need to discern sarcasm and other nuanced sentiments.

Meanwhile, the Dataset for Emotion Classification (DEC) was curated from licensed digital literary databases, ensuring respect for intellectual property. Extraneous details unrelated to the emotional tone were removed. The dataset categorized emotions into 8,000 samples for joy (20%), 7,000 for sadness (17.5%), 7,500 for anger (18.75%), 6,500 for fear (16.25%), 6,000 for surprise (15%), and 5,000 neutral samples (12.5%).

For the Indonesian Dataset for Tweet Sentiment Analysis (IDTSA) and the Minangkabau Dataset for Tweet Sentiment Analysis (MDTSA), the sentiment distribution yielded 7,359 positive tweets (57%), 1,367 neutral (11%), and 4,034 negative (32%). Meanwhile, the Indonesian Dataset for Emotion Classification (IDEC) and the Minangkabau Dataset for Emotion Classification (MDEC) categorized emotions into 637 samples for love (14.47%), 1,017 samples for happiness(23.10%), 998 samples for sadness (22.66%), 1,102 samples for anger (25.03%), and 649 samples for fear (14.74%).

In summary, this study utilized a variety of datasets specifically curated for different natural language processing tasks in Turkish, Indonesian, and Minangkabau. Below, we provide a detailed description of these datasets, outlining their sources, sizes, and characteristics to offer insights into the data foundation of our research.

a: DATASETS FOR TURKISH LANGUAGE TASKS

- **Dataset for Topic Classification (DTC):** Comprising 60,000 samples sourced from proprietary collections and various Turkish newspapers. This dataset is a mix of genres including news articles, academic journals, and literature, which are categorized into themes such as Current Affairs, Economics, Technology, and more. The diverse sources ensure a broad coverage of topics and linguistic styles.
- **Dataset for Tweet Sentiment Analysis (DTSA):** This dataset includes 50,000 tweets collected via the Twitter API, labeled for sentiment analysis with categories including Positive, Negative, and Neutral. The dataset is designed to reflect a wide range of public opinions and emotional tones.
- **Dataset for Emotion Classification (DEC):** Contains 40,000 excerpts from contemporary Turkish literature, classified into emotions like Happiness, Sadness, Anger, Surprise, and Fear, providing rich linguistic expressions of diverse emotional states.

b: DATASETS FOR INDONESIAN LANGUAGE TASKS

- **Indonesian Dataset for Tweet Sentiment Analysis (IDTSA):** Consisting of comments and reviews in Indonesian, used for sentence-level sentiment analysis with categories such as Positive, Negative, and Neutral. This dataset helps in understanding the sentiment distribution in consumer feedback and social media interactions.
- **Indonesian Dataset for Emotion Classification (IDEC):** Includes 4,403 tweets sourced from Twitter, categorized into emotions including Love, Joy, Surprise, Anger, Sadness, and Fear. It is useful for studying the emotional content in brief social media texts.

c: DATASETS FOR MINANGKABAU LANGUAGE TASKS

- **Minangkabau Dataset for Tweet Sentiment Analysis (MDTSA)** and **Minangkabau Dataset for Emotion Classification (MDEC):** These datasets are translations of the IDTSA and IDEC, respectively, using a bilingual Indonesian-Minangkabau dictionary. They are tailored for sentiment and emotion classification in the Minangkabau language, facilitating the study of this less-resourced linguistic context.

These datasets were meticulously selected and curated to ensure a comprehensive analysis across various NLP tasks, providing a robust foundation for evaluating the performance of both human annotators and Large Language Models in processing low-resource languages.

B. NLP TASKS AND ANNOTATION GUIDELINES

Each of our chosen NLP tasks poses its unique challenges, demanding comprehensive guidelines to ensure clarity, consistency, and accuracy. In this section, we expand on the task intricacies and offer detailed guidelines, mindful of the nuances of the Turkish, Indonesian, and Minangkabau language.

- **Topic Classification** aims to allocate textual information to distinct predefined topics. The intrinsic richness of the Turkish language, characterized by a plethora of synonyms and its dynamic range of expression, amplifies the intricacies of this task, especially within

the realms of academic and journalistic content. Annotators were guided to perform a holistic contextual analysis, acknowledging the polysemic nature of many Turkish terms. For instance, topics like 'Current Affairs' could be demarcated by time-bound terms such as 'son dakika' (latest news), while 'Geopolitics' might encompass terminology related to international affairs, treaties, or words like 'diplomasi' (diplomacy). In mitigating biases, particularly pertinent in polarized news content, the emphasis was laid on the textual narrative, overshadowing potential media leanings. Indonesian language, an artificial language created to unify Indonesian people is greatly influenced by Indonesian ethnic languages such as Javanese, Minangkabau, Buginese and Banjarese. Moreover, many borrowed words from Arabic, Dutch, and English have been adapted to fit the phonetic and grammatical rules of Indonesian language. This contributes to the polysemic nature of many Indonesian words such as 'akar' which has two meanings: 'root' as in the root of a tree and 'source' as in the source of a problem, and the homonymic nature such as 'tinggi' which has two meanings: 'high/elevated' refers to the measurement of height or elevation and 'loud' refers to a high volume or intensity of sound. The polysemous words requires special attention as one of critical and identified problem of natural language processing [33], [34], [35], [36]. Minangkabau, an Austronesian language specifically belonging to the Malayic languages spoken in West Sumatra, Indonesia, is known by ChatGPT-4, but not listed as one of 32 languages that ChatGPT-4 understand well like Indonesian and Turkish. Minangkabau and Indonesian are distinct languages but share a historical and geographical relationship due to Indonesia's diverse linguistic landscape. The relationship between Minangkabau and Indonesian is influenced by historical, cultural, and linguistic factors. Many Minangkabau speakers are also fluent in Indonesian, which is taught in schools and used as a lingua franca across the archipelago. Minangkabau terms have a polysemic nature similar to Indonesian terms. Therefore, topic classification poses a similar challenge for Minangkabau as it does for Indonesian.

- **Sentiment analysis** within the context of tweets is inherently challenged by the platform's colloquialisms, condensation, and culture-specific digital expressions. The brevity, teamed with the frequent deployment of Turkish, Indonesian, and Minangkabau specific emojis and internet slang, accentuates the task's complexity. Annotation strategies were anchored around local linguistic markers—terms like 'aşkım' (my love) or 'yok artık' (can't believe it) which possess potent sentiment indicators in Turkish. Additionally, certain emojis, might universally evoke autumnal sentiments, hold deeper melancholic or nostalgic undertones within the Turkish digital discourse. Keeping annotators attuned to contemporary events was crucial, ensuring accurate interpretation of tweets with references to the current zeitgeist.
- Diving into the realm of literature for **Emotion Classification** necessitates a profound understanding of Turkey's cultural, historical, and literary fabric. The intricate tapestry of literary narratives, deeply entangled with societal constructs, demands a discerning eye for accurate emotional categorization. Annotation guidelines highlighted the importance of cultural cues, such as references to iconic figures like 'Karagöz and Hacivat', indicative of humor or satire. Annotators were trained to identify and decode prevalent literary instruments in Turkish literature, such as metaphors and allegories, to unveil embedded emotions. In scenarios of ambiguity or multifaceted emotional representation, collaborative cross-referencing and supplementary sources were harnessed for validation. The aforementioned guidelines can be utilized for both Indonesian and Minangkabau annotators as well.

1) TURKISH ANNOTATION GUIDELINES

- **Lexical Ambiguity:** Annotators were trained to identify and resolve ambiguous terms based on context, using surrounding text and external resources to discern the correct meanings of polysemous words.
- **Cultural References:** Comprehensive instructions were included on cultural idioms or phrases unique to Turkish culture to ensure correct understanding and classification.
- **Syntax and Grammar:** Attention was given to Turkish syntax that might affect meaning, such as negation or the placement of adjectives.
- **Training and Calibration:** Annotators underwent preliminary training sessions with examples of annotated texts. Regular calibration meetings were held to discuss challenging cases and ensure consistent application of guidelines.

2) INDONESIAN ANNOTATION GUIDELINES

- **Handling of Formal and Informal Language:** Annotators were provided with examples of formal (Bahasa Indonesia) and informal (slang) usages and trained to identify and differentiate between them in text.
- **Polysemy and Homophony:** Guidelines were developed to instruct annotators on the context-dependent meanings of words, with specific examples provided for clarity.
- **Sentiment Indicators:** Special training was provided for identifying sentiment in Indonesian, considering modifiers and intensifiers that significantly affect sentiment expression.
- **Consistency Checks:** Double annotation was employed, where two annotators independently labeled the same text, followed by a discussion to resolve any discrepancies.

3) MINANGKABAU ANNOTATION GUIDELINES

- **Dialectal Variations:** Annotators were made aware of key dialectical variations within Minangkabau and trained on regional linguistic features that might influence text interpretation.
- **Cultural Expressions:** Instructions included how to interpret phrases and expressions tied to local customs, which are pivotal for accurate emotion classification.
- **Emotional Tone Recognition:** Given the importance of correctly identifying emotional undertones in Minangkabau, annotators were trained with specific emotional categories and examples illustrating each.
- **Review and Adjustment Process:** Annotators were required to periodically review their annotations with a senior linguist to discuss and refine their interpretations based on feedback.

4) TURKISH LINGUISTIC NUANCES

Turkish is notable for its agglutinative nature, which allows the creation of complex sentences from a single verb root through various affixes. For instance, the verb *görmek* (to see) can transform into *görüyorum* (I see/am seeing), *görüyordum* (I was seeing), and *görebilirim* (I might see). This poses challenges in semantic parsing, where the precise meaning changes subtly with different suffixes.

Moreover, the use of evidentiality in Turkish—a grammatical mood indicating the source of information—is exemplified by the suffix *-miş*. For example, *geldi* means "he/she came (and I saw it)," whereas *gelmiş* means "he/she has come (I heard/assumed)." Evidentiality requires NLP systems to understand nuances beyond direct observation, impacting tasks like sentiment analysis and fact-checking.

5) INDONESIAN LINGUISTIC CHARACTERISTICS

Indonesian's use of affixes profoundly impacts its syntax and semantics. The prefix *pem-* and suffix *-an* can change a verb to a noun indicating a collective or abstract form, such as from *buka* (open) to *pembukaan* (opening ceremony). Handling these transformations is essential for accurate machine translation and information extraction.

Another complexity is the informal versus formal usage, illustrated by the use of *kau* (you, informal) and *Anda* (you, formal). The choice between these can alter the tone and social context understood by AI in interactive applications like chatbots.

6) MINANGKABAU LINGUISTIC CHARACTERISTICS

Minangkabau, with its rich oral tradition, includes idiomatic expressions that are deeply cultural. For instance, *alam takambang jadi guru* (the nature becomes the teacher) is an idiom used to express the value of life experiences, posing challenges for translation and cultural sensitivity in NLP applications.

The language's morphology also reflects its Austronesian roots, where prefixes and suffixes are used extensively, similar to Indonesian. For example, the addition of *di-* (passive marker) and *-kan* (causative suffix) to the root *tulis* (write) forming *dituliskan* (is written for/by someone), requires detailed syntactic and contextual analysis for correct interpretation in NLP systems.

C. DATASET SELECTION AND CURATION

The datasets for this study were carefully selected to represent a diverse spectrum of text types within each of the target low-resource languages: Turkish, Indonesian, and Minangkabau. The selection process was guided by the following criteria:

- **Language Representation:** We ensured that each dataset contained a balanced representation of various linguistic features, including idiomatic expressions, colloquial language, and formal text.
- **Text Diversity:** Texts included a mix of genres such as news articles, literary works, and informal online communications to mimic the variety of real-world applications.
- **Data Availability and Accessibility:** Only publicly available or ethically sourced data were used to construct the datasets, adhering to all relevant data protection regulations.
- **Annotation Quality:** Pre-annotated datasets underwent a rigorous quality check to ensure that existing annotations met high standards of accuracy and relevance.

The curation process involved preprocessing techniques to standardize text formats, remove noise, and anonymize personal information, thereby preparing the data for effective use in training and evaluating NLP models.

D. EVALUATION CRITERIA FOR LLMS AND HUMAN ANNOTATORS

The performance of both LLMs and human annotators was evaluated using the following metrics, which are standard in the field of natural language processing:

- **Precision:** This metric measures the accuracy of the annotations provided, defined as the proportion of true positive results divided by the number of all positive results reported by the classifier.
- **Recall:** Recall assesses the completeness of the annotations, defined as the proportion of true positive results divided by the number of positives that should have been retrieved.
- **F1-Score:** The F1-Score is the harmonic mean of precision and recall, providing a single metric to evaluate the balance between precision and recall.
- **Inter-Annotator Agreement:** For human annotators, the consistency of annotation was evaluated using Kappa statistics, which measure the agreement between annotators beyond chance.

These metrics were chosen for their ability to provide a comprehensive assessment of performance across different types of NLP tasks and to allow for direct comparison between human and machine-generated annotations.

E. HUMAN ANNOTATORS

In this section, we provide a detailed account of the human annotation process employed in this study. Human annotation played a pivotal role in ensuring the quality and reliability of the labeled data used for our Natural Language Processing (NLP) tasks. The process encompassed annotator selection, training, the formulation of annotation guidelines, the choice of annotation platform, assessment of inter-annotator agreement, and strategies to address encountered challenges. The selection of human annotators was a critical step in our annotation process. Annotators were carefully chosen based on their proficiency in the language and their familiarity with the specific NLP tasks under consideration. We provided training to three students to perform the annotation tasks. We sought annotators with a strong linguistic background and domain knowledge relevant to the topics covered in our datasets. This selection criterion ensured that the annotators were well-equipped to provide accurate annotations in alignment with the objectives of our research. Prior to embarking on the annotation tasks, annotators underwent a rigorous training program tailored to each NLP task. Training sessions were designed to acquaint annotators with the annotation guidelines, task-specific challenges, and the intricacies of the language as it related to the tasks. These training sessions included the presentation of task-specific examples and encouraged annotators to seek clarifications and ask questions to establish a clear understanding of the guidelines. The creation of comprehensive and task-specific annotation guidelines was central to maintaining consistency and accuracy in the human annotation process. These guidelines provided annotators with a standardized approach for labeling the data. They included explicit definitions of task-specific labels, instructions for handling ambiguous cases, and guidance on addressing context-dependent linguistic nuances. The guidelines were continuously refined and adapted to address the unique challenges posed by each NLP task, which included topic classification, tweet sentiment analysis, and emotion classification. To facilitate the efficient and consistent annotation of our datasets, a custom annotation platform was developed. This platform enabled annotators to access the data, apply annotations in accordance with the provided guidelines, and submit their annotations in a structured format. Additionally, the platform supported real-time communication between annotators and the research team, allowing for prompt query resolution and ensuring the quality of annotations.

To evaluate the reliability and consistency of the human annotations, we conducted inter-annotator agreement assessments. A subset of the data was independently annotated by multiple annotators, and the level of agreement was quantified using standard metrics. This iterative process of cross-annotation helped ensure that the annotations were reliable and that annotators were aligned in their interpretations of the guidelines.

Throughout the human annotation process, various challenges were encountered, such as interpreting context-dependent expressions, handling colloquial language in tweets, and addressing ambiguities in literary texts. To mitigate these challenges, regular meetings were held with annotators to address questions, provide clarifications, and offer guidance. An iterative feedback loop was established to refine the annotation guidelines and enhance the overall quality of annotations over the course of the project.

F. MECHANICAL TURK ANNOTATORS

This section provides an in-depth exploration of the annotation methodology using Amazon Mechanical Turk (MTurk) as a critical component of our research. We enlisted the assistance of MTurk workers for carrying out the identical set of tasks as our trained annotators and ChatGPT, using the same set of guidelines. To maintain the quality of annotations, we limited task access to workers designated as ''MTurk Masters'' by Amazon. These workers were required to have a HIT (Human Intelligence Task) approval rate exceeding 90% with at least 50 approved HITs [20].

MTurk annotation offers distinct advantages and considerations, contributing to the comprehensive data collection strategy employed in this study. Amazon Mechanical Turk (MTurk) serves as a versatile platform for large-scale annotation tasks, facilitating the efficient processing of substantial volumes of data. Key aspects include:

- MTurk allows for the swift expansion of annotation efforts, making it well-suited for projects requiring annotations on a massive scale.
- **Global Workforce**: MTurk provides access to a diverse and global workforce, offering a broad range of perspectives and linguistic capabilities.
- **Cost-Efficiency**: For tasks that do not necessitate specialized domain knowledge, MTurk can be a cost-effective option, making it particularly suitable for projects with budget constraints.

One of the notable strengths of MTurk annotation lies in its ability to harness a diverse pool of workers from various backgrounds. This diversity contributes to a richer dataset by incorporating multiple viewpoints and language proficiencies:

Cultural and Linguistic Diversity: MTurk workers represent a wide range of cultures and languages, enhancing the dataset's diversity and inclusivity. Varied Expertise: MTurk workers may bring diverse expertise to the task, which can be valuable for certain annotation projects requiring specific knowledge. While MTurk annotation offers scalability and diversity, maintaining annotation quality is paramount. Several quality control mechanisms were implemented to ensure the reliability of MTurk-generated annotations: Multiple annotations for the same data points were collected to assess agreement among MTurk workers. An additional validation step involved cross-checking annotations for consistency and accuracy, allowing for the identification and rectification of discrepancies. Feedback loops were established to provide MTurk workers with guidance and clarifications, contributing to improved annotation quality. It is essential to

acknowledge that the suitability of MTurk annotation varies depending on the complexity and specificity of the task: MTurk annotation is highly efficient for straightforward tasks but may face challenges with complex or domain-specific assignments. In cases where domain expertise is required, MTurk annotation may need to be supplemented with input from domain specialists to ensure accurate annotations. In our research, MTurk annotation played a pivotal role in tasks where scalability and diverse perspectives were essential.

G. METRICS FOR ANNOTATION QUALITY EVALUATION

This section provides a comprehensive understanding of the metrics and criteria used to evaluate the quality of annotations generated through different methodologies, including ChatGPT, Mechanical Turk (MTurk), and human annotation. These metrics play a crucial role in assessing the effectiveness and reliability of each annotation approach.

Precision is a fundamental metric that measures the accuracy of annotations. It answers the question: ''Of all the instances that the methodology labeled as belonging to a specific category, how many were truly relevant?'' Precision is crucial in ensuring that the annotations correctly identify the intended category without including unrelated instances. Precision is calculated as:

$$Precision = TP/(TP + FP) \tag{1}$$

where TP denotes instances correctly identified by the methodology as belonging to the specified category, and FP denotes instances incorrectly labeled as belonging to the specified category.

Recall assesses the completeness of annotations. It answers the question: ''Of all the instances that should have been labeled as belonging to a specific category, how many were correctly identified by the methodology?'' Recall is crucial for ensuring that no relevant instances are missed during annotation. Recall is calculated as:

$$Recall = TP/(TP + FN) \tag{2}$$

where FN denotes instances that should have been labeled as belonging to the specified category but were missed.

The F1-score is a combined metric that balances precision and recall. It provides a comprehensive assessment of annotation quality, particularly in situations where both precision and recall are essential. A high F1-score indicates that the annotations are both accurate (high precision) and comprehensive (high recall).

Inter-annotator agreement evaluates the consistency and reliability of annotations generated by multiple human annotators. In our research, we employ Fleiss' Kappa coefficient, a well-established measure of agreement. A higher Kappa coefficient indicates a greater level of agreement among human annotators, which is desirable for reliable annotations.

H. LARGE LANGUAGE MODELS

In this section, we introduce the Large Language Models (LLMs) that serve as the cornerstone of our annotation generation process. These models play a pivotal role in facilitating the production of annotations and enabling a comprehensive comparative analysis of annotation quality.

ChatGPT-4, an exemplary creation by OpenAI, represents the pinnacle of large-scale language models [37], [38]. With a staggering 175 billion parameters, this model exhibits an unparalleled understanding of human language and context. ChatGPT-4's prowess lies in its ability to generate coherent, contextually aware text, making it an ideal candidate for generating annotations across a spectrum of Natural Language Processing (NLP) tasks.

In parallel with our exploration of ChatGPT and its variants, we also considered the inclusion of other prominent LLMs to diversify our approach to annotation generation. These models include:

BERT (Bidirectional Encoder Representations from Transformers): BERT, known for its bidirectional contextual understanding, has established itself as a formidable entity in the realm of LMs [39]. Its architecture, grounded in the Transformer framework, allows it to capture nuanced language nuances effectively. BERTurk is fine-tuned specifically for the Turkish language, making it well-suited for various natural language processing (NLP) tasks in Turkish. Similar to other BERT variants, it can be used for tasks such as text classification, sentiment analysis, named entity recognition, and more. When comparing annotation schemes, we also consider BERTurk as one of the LLMs for generating annotations to assess its performance in Turkish NLP tasks.

RoBERTa (A Robustly Optimized BERT Pretraining Approach): RoBERTa, a refinement of the BERT architecture, places a strong emphasis on robust optimization techniques [40]. This emphasis results in heightened language understanding capabilities, allowing RoBERTa to excel in generating annotations with a nuanced touch.

T5 (Text-To-Text Transfer Transformer): T5, underpinned by the Text-To-Text Transfer Transformer framework, presents an innovative approach to language understanding [41]. Its text-to-text methodology enables versatile applications across language tasks, positioning it as a noteworthy contender for annotation generation.

ChatGPT-4's architectural magnificence lies in its multi-layered neural network structure, underpinned by transformers. These transformers enable ChatGPT-4 to navigate the complexities of context, syntax, and semantics with extraordinary precision. This profound understanding of language forms the bedrock of its annotation generation capabilities.

In tandem with our exploration of ChatGPT and its variants, we harnessed the capabilities of other eminent LLMs, including BERT, RoBERTa, and T5, for the purpose of annotation generation. While these models were not subjected to fine-tuning for the Turkish language, they were rigorously evaluated for their performance in generating annotations across the same set of Turkish NLP tasks. This comprehensive evaluation affords us valuable insights and forms the basis for our ensuing comparative analysis.

The process of generating annotations using LLMs involved the following steps:

- **Data Preparation**: We provided the LLMs with carefully curated datasets specific to each NLP task - topic classification, tweet sentiment analysis, and emotion classification. These datasets consisted of Turkish text samples relevant to the respective tasks.
- **Input Prompt**: For each annotation task, we formulated input prompts that were specific to the nature of the task. These prompts served as instructions to guide the LLMs in generating annotations. For example, for topic classification, the input prompt may instruct the LLM to categorize a given text sample into predefined topics.
- **Annotation Generation**: The LLMs, including ChatGPT-4, BERT, RoBERTa, and T5, were tasked with generating annotations based on the input prompts. They utilized their pre-trained language understanding to produce annotations that included classifications, sentiment labels, or emotional categorizations as required by the respective NLP tasks.
- **Quality Control**: To ensure the quality of the generated annotations, we implemented rigorous quality control measures. This included assessing the coherence, relevance, and accuracy of the annotations produced by the LLMs.

IV. EXPERIMENTAL SETTINGS AND RESULTS

In this section, the experimental procedure and the results obtained by the compared schemes have been presented.

A. EXPERIMENTAL SETTINGS

In this section, we detail the essential components of our experimental setup, encompassing the dataset split, annotation tools, and Large Language Model (LLM) configuration:

- **Training, Validation, and Test Sets**: The dataset division was executed with meticulous care to ensure robust evaluation and reliable results. We adopted a standard split methodology for each of our three NLP tasks: topic classification, tweet sentiment analysis, and emotion classification.
- **Training Set**: The training set, comprising a significant portion of the dataset, was utilized for training the machine learning models. It constituted 70% of the total dataset for each NLP task. The training set was instrumental in allowing our models to learn the underlying patterns, semantics, and characteristics of the respective tasks.
- **Validation Set**: The validation set, constituting 15% of the dataset, played a pivotal role in model fine-tuning and parameter optimization. It facilitated the selection of optimal hyperparameters and the prevention of overfitting.
- **Test Set**: The remaining 15% of the dataset was allocated to the test set. This independent test set was kept separate from the training and validation data and was employed for the final evaluation of model performance. It allowed us to assess the generalization capabilities of our models on unseen data.

To generate annotations for our NLP tasks, we employed a combination of annotation methodologies:

- **Human Annotation**: Human annotators, proficient in the Turkish language, were tasked with producing high-quality annotations for each NLP task. They adhered to comprehensive annotation guidelines to ensure consistency and accuracy in the annotations.
- **Mechanical Turk (MTurk)**: For comparative analysis, we utilized the Mechanical Turk platform to obtain annotations from crowd-workers. Crowd-workers followed the same annotation guidelines as human annotators.
- **LLM Annotation**: Large Language Models (LLMs), including ChatGPT-4, BERT, RoBERTa, and T5, were employed to generate annotations. These models were guided by task-specific input prompts and fine-tuned for linguistic and cultural relevance, as described in the previous section.

In this study, feature extraction from the text was a critical step for enabling the NLP models to perform various tasks such as sentiment analysis, emotion classification, and topic classification. Initially, raw text data from diverse sources was preprocessed to remove irrelevant content and normalize the text. Subsequently, feature extraction techniques were employed to transform the text into a format that could be understood by machine learning models. In our study, we implement a comprehensive feature extraction framework to preprocess and transform text data for optimal performance with NLP models. The process begins with rigorous text preprocessing, involving normalization, tokenization, and noise reduction to ensure a clean dataset. We then employ advanced vectorization techniques, such as contextual embeddings from transformer-based models, which capture deep linguistic features far beyond traditional Bag-of-Words or TF-IDF methods. A critical component of our approach is the integration of syntactic and semantic analysis. By leveraging dependency parsing and named entity recognition, we extract rich syntactic and semantic features that contribute significantly to the understanding of text context and structure. Additionally, we explore the use of custom feature engineering, tailored specifically to the linguistic characteristics of our target languages, enhancing the models' ability to handle language-specific nuances. To address the high-dimensionality of our feature vectors, dimensionality reduction techniques like t-SNE and UMAP are applied. This not only improves model efficiency but also aids in uncovering underlying patterns within the text data. Our feature extraction methodology is designed to be robust and adaptable, capable of being applied across various NLP tasks, from sentiment analysis to more complex language understanding and generation tasks. For languages like Turkish, Indonesian, and Minangkabau, which have limited language resources, the extraction process involved leveraging the syntactic and semantic properties of the text.

The models employed, including ChatGPT-4, BERT, and others, utilized their respective tokenization methods to decompose text into meaningful units. These units, or tokens, were then vectorized to represent linguistic features like word context, semantic meaning, and part-of-speech tags. The vectorized features served as the input for the NLP models, allowing them to perform the designated tasks with a higher degree of accuracy. The robustness of feature extraction was essential for ensuring that the models could effectively interpret the nuances of these languages, which often contain complex linguistic structures. Overall, the approach to feature extraction in this study was tailored to address the challenges posed by languages with limited digital resources, demonstrating the versatility and adaptability of advanced NLP models in diverse linguistic contexts. In our research, the parameterization of the Large Language Models (LLMs) such as ChatGPT-4 and BERT was a pivotal aspect. To effectively tailor these models for our specific NLP tasks, we conducted a series of preliminary experiments aimed at identifying the optimal configuration for each model. The determination of parameters involved several considerations: Model Size and Complexity: We evaluated different sizes of the models (measured in the number of parameters) to strike a balance between computational efficiency and performance accuracy. Training Data: The scope and diversity of the training data were critically analyzed to ensure that the models were well-equipped to understand and process the linguistic intricacies of Turkish, Indonesian, and Minangkabau languages. Fine-Tuning: We implemented fine-tuning processes, where the models were further trained on task-specific datasets to enhance their accuracy in annotation generation and other NLP tasks Hyperparameter Optimization: Bayesian optimization was employed to systematically explore various hyperparameter configurations, ensuring the selection of the most effective settings for our models. Evaluation Metrics: The models' parameters were also influenced by the desired performance metrics, such as precision, recall, and F1-score, which guided our parameter tuning process to align with the objectives of our NLP tasks. By meticulously determining the parameters of the LLMs, our study ensured that the models were not only theoretically sound but also practically effective in handling the unique challenges posed by the selected languages and tasks. Our experimental setup included a diverse range of LLMs, each contributing to the annotation generation process. Here, we provide an overview of the key LLMs and their configurations:

ChatGPT-4: ChatGPT-4 is a colossal LLM with 175 billion parameters, renowned for its ability to produce coherent and contextually relevant text. In addition to ChatGPT, we incorporated other LLMs, including BERT, BERTURK, RoBERTa, and T5, into our annotation generation process. While some of these models were not fine-tuned for Turkish, they were assessed for their performance in generating annotations across the same set of Turkish NLP tasks, contributing to our comparative analysis.

Large Language Models (LLMs) Configuration: For this study, we utilized several state-of-the-art Large Language Models to generate annotations for comparison with human-generated data. The specific models included:

BERT (Bidirectional Encoder Representations from Transformers) - Configured for the Turkish language (BERTurk), this model has been extensively fine-tuned on a diverse corpus of Turkish text, enabling it to understand and process Turkish syntactic and semantic nuances effectively.

RoBERTa (A Robustly Optimized BERT Pretraining Approach) - Utilized for its robust performance on text classification tasks, RoBERTa was configured with default pretraining followed by fine-tuning on our specific datasets to align with the linguistic challenges presented by the Turkish, Indonesian, and Minangkabau languages.

GPT-3 (Generative Pre-trained Transformer 3) - Deployed for its advanced generative capabilities, GPT-3 was fine-tuned on a mixture of publicly available and proprietary datasets that mirror the linguistic diversity and complexity of our target NLP tasks.

Each model was run with the following settings:

Batch size: 32 Learning rate: 5e-5, with a linear decay schedule applied over the training epochs Epochs: 4 for fine-tuning, ensuring that the models do not overfit to the training data. Our training data comprised a balanced set of annotations across different text types to ensure that the model training phases covered a comprehensive range of linguistic structures and colloquial uses specific to each language.

B. EXPERIMENTAL RESULTS

In order to establish a robust gold standard for evaluating the quality of annotations in our research, we adopted a comprehensive scheme that considers both average accuracy and intercoder agreement. This gold-standard creation process was instrumental in ensuring the reliability and validity of our annotation evaluation.

We initiated the gold-standard creation process by computing the average accuracy of annotations. Average accuracy represents the percentage of correct predictions made by human annotators. To compute this metric, we selectively considered texts where there was a unanimous agreement between the two annotators. This means that we only included instances for which both annotators independently assigned the same class label. In other words, we focused on instances where there was a consensus in the annotations provided by human experts.

In addition to average accuracy, we assessed intercoder agreement, which measures the extent to which annotators within the same group report the same class label for a given instance. To compute this metric, we examined the percentage of instances where both annotators assigned identical class labels independently. This measure enabled us to evaluate the degree of concordance between annotators and assess their consistency in classification.

In Table 2, we present the precision scores obtained for the various annotation schemes across different datasets and classes. Precision, as a fundamental metric, measures the accuracy of positive predictions made by each annotation scheme. It is a crucial indicator of the annotation quality, reflecting how well each scheme correctly identifies instances belonging to a specific class. The DTC (Dataset for Topic Classification) encompasses a wide range of topics, making it a challenging dataset for annotation. Human annotations exhibit high precision across most classes, indicating the accuracy of human experts in classifying topics. Notably, the ''Humanities'' and ''Natural Sciences'' categories stand out with precision scores of 0.973 and 0.958, respectively. This suggests that human annotators excelled in accurately categorizing texts within these domains. Among the Language Models (LLMs), ''BERTurk'' demonstrates competitive precision, especially in the ''Economics'' and ''Societal Issues'' categories.

In Table 3, we delve into the recall scores obtained for the various annotation schemes across different datasets and classes. Recall measures the ability of each annotation scheme to correctly identify positive instances within each class. We will also draw comparisons with the precision values discussed earlier to provide a comprehensive view of the annotation quality. The recall scores for the DTC (Dataset for Topic Classification) reveal notable trends. Human annotations maintain high recall values, reflecting their effectiveness in identifying instances across diverse topics. Precision-recall trade-offs are evident as certain LLMs, such as ''BERT'' and ''RoBERTa,'' demonstrate improved recall values compared to their precision. ''ChatGPT-4'' and ''BERTurk'' maintain competitive recall scores, particularly in categories like ''Current Affairs'' and ''Economics.''

The recall scores for the DTSA (Dataset for Tweet Sentiment Analysis) dataset illustrate the performance of annotation schemes in sentiment analysis. Human annotations continue to exhibit strong recall, aligning with their precision scores. Notably, ''Positive'' sentiments achieve higher recall values compared to other sentiments, highlighting the proficiency of human annotators in identifying positivity in tweets. Among LLMs, ''BERT'' showcases balanced recall scores across sentiments.

In the DEC (Dataset for Emotion Classification), recall scores portray the ability of annotation schemes to classify text excerpts into emotion categories. Human annotations exhibit variations in recall across emotions, with ''Sadness'' achieving the highest recall at 0.914. LLMs like ''BERT'' and ''BERTurk'' demonstrate competitive recall values, indicating their capacity to identify emotions effectively.

Comparing recall scores with precision values, we observe precision-recall trade-offs. Human annotations consistently maintain high precision and recall, demonstrating their proficiency in various NLP tasks. LLMs, while competitive, often exhibit differences in precision-recall balances. For example, ''ChatGPT-4'' and ''BERTurk'' demonstrate better recall in certain classes but slightly lower precision compared to human annotations. The observed precision-recall trade-offs underscore the complexities of annotation quality. While LLMs show promise in specific contexts, they may prioritize recall over precision, highlighting the challenges in handling nuanced language and context.

In Table 4, we present the F1-scores obtained for various annotation schemes across different datasets and classes. F1-score, the harmonic mean of precision and recall, provides a balanced measure of annotation quality. We will also compare these scores to the precision and recall values discussed earlier to gain a comprehensive understanding of annotation quality. The F1-scores for the DTC (Dataset for Topic Classification) dataset offer insights into annotation quality across diverse topics. Human annotations consistently exhibit high F1-scores, indicating their proficiency in classifying texts. Precision-recall trade-offs are evident as some LLMs, such as ''ChatGPT-4'' and ''BERTurk,'' achieve improved F1-scores compared to precision alone. These trade-offs reflect the challenges in balancing precision and recall. F1-scores for the DTSA (Dataset for Tweet Sentiment Analysis) dataset provide a perspective on sentiment analysis. Human annotations maintain strong F1-scores across sentiments, demonstrating their ability to identify sentiment effectively. Among LLMs, ''BERT'' consistently achieves balanced F1-scores across sentiment categories, emphasizing its proficiency in sentiment analysis tasks.

The F1-scores for the DEC (Dataset for Emotion Classification) dataset shed light on the classification of text excerpts into emotion categories. Human annotations exhibit variations in F1-scores across emotions, with ''Sadness'' achieving the highest F1-score at 0.922. LLMs, such as ''BERT'' and ''BERTurk,'' demonstrate competitive F1-scores, indicating their capacity to classify emotions effectively. Comparing F1-scores with precision and recall values, we observe the balance achieved by each annotation scheme. Human annotations consistently demonstrate high precision, recall, and F1-scores across datasets and classes, signifying their reliability. LLMs, while competitive, often exhibit variations in the balance between precision, recall, and F1-score. For instance, ''BERT'' showcases balanced F1-scores across sentiments and emotions. The observed balance between precision, recall, and F1-score highlights the complexities of annotation quality assessment. While LLMs offer promise in specific contexts, their performance may vary across different NLP tasks and classes. In summary, the F1-scores provide a balanced perspective on annotation scheme performance, reflecting precision-recall trade-offs and variations. These findings contribute to a comprehensive understanding of annotation quality in Turkish NLP tasks.

In Figure 1, the main effects plot for precision values of annotation schemes has been presented. The Main Effects Plot illuminates precision disparities across multiple models and annotation strategies. Starting with BERT's foundational precision slightly above 0.82, BERTurk shows a subtle improvement, suggesting enhanced capabilities. ChatGPT-4's precision closely matches that of Human

TABLE 2. The precision values for the compared annotation schemes.

Annotation Scheme	Human Annotation	MTurk Annotation	ChatGPT-4	BERT	BERTurk	RoBERTa	T5
DTC Dataset	0.913	0.773	0.879	0.843	0.877	0.789	0.814
Current Affairs	0.941	0.819	0.842	0.799	0.876	0.846	0.822
Geopolitics	0.908	0.745	0.864	0.827	0.865	0.810	0.763
Economics	0.873	0.767	0.872	0.828	0.955	0.778	0.899
Societal Issues	0.902	0.696	0.892	0.911	0.839	0.780	0.873
Humanities	0.973	0.790	0.913	0.871	0.867	0.801	0.849
Social Sciences	0.867	0.824	0.856	0.788	0.875	0.670	0.780
Natural Sciences	0.958	0.793	0.910	0.801	0.810	0.771	0.798
Prose	0.884	0.767	0.839	0.892	0.908	0.791	0.813
Poetry	0.913	0.753	0.925	0.868	0.900	0.851	0.724
DTSA Dataset	0.875	0.781	0.906	0.809	0.836	0.791	0.799
Positive	0.933	0.736	0.912	0.829	0.758	0.785	0.735
Negative	0.840	0.888	0.870	0.772	0.925	0.764	0.836
Neutral	0.852	0.717	0.937	0.826	0.824	0.825	0.827
DEC Dataset	0.857	0.779	0.868	0.824	0.894	0.823	0.761
Happiness	0.898	0.727	0.817	0.769	0.862	0.898	0.746
Sadness	0.930	0.839	0.895	0.765	0.894	0.758	0.747
Anger	0.868	0.799	0.946	0.809	0.926	0.800	0.853
Fear	0.778	0.788	0.941	0.812	0.811	0.774	0.700
Surprise	0.911	0.763	0.753	0.916	0.970	0.818	0.719
Disgust	0.862	0.753	0.851	0.797	0.869	0.828	0.782
Neutral	0.754	0.785	0.870	0.901	0.925	0.888	0.784

TABLE 3. The recall values for the compared annotation schemes.

Annotation Scheme	Human Annotation	MTurk Annotation	ChatGPT-4	BERT	BERTurk	RoBERTa	T5
DTC Dataset	0.883	0.759	0.882	0.876	0.860	0.804	0.787
Current Affairs	0.962	0.678	0.936	0.911	0.857	0.782	0.803
Geopolitics	0.830	0.796	0.905	0.947	0.825	0.785	0.784
Economics	0.858	0.798	0.839	0.829	0.839	0.843	0.748
Societal Issues	0.871	0.741	0.826	0.806	0.844	0.799	0.762
Humanities	0.859	0.794	0.954	0.876	0.871	0.835	0.717
Social Sciences	0.868	0.743	0.883	0.933	0.820	0.824	0.734
Natural Sciences	0.917	0.797	0.868	0.743	0.920	0.725	0.829
Prose	0.862	0.725	0.862	0.950	0.940	0.823	0.867
Poetry	0.925	0.762	0.862	0.885	0.827	0.821	0.839
DTSA Dataset	0.891	0.782	0.917	0.805	0.909	0.815	0.844
Positive	0.937	0.779	0.941	0.858	0.936	0.736	0.790
Negative	0.882	0.771	0.912	0.790	0.884	0.867	0.848
Neutral	0.855	0.795	0.899	0.767	0.907	0.841	0.895
DEC Dataset	0.891	0.761	0.903	0.859	0.871	0.805	0.797
Happiness	0.849	0.755	0.941	0.886	0.925	0.807	0.824
Sadness	0.914	0.773	0.925	0.871	0.815	0.787	0.828
Anger	0.890	0.726	0.894	0.802	0.905	0.817	0.748
Fear	0.855	0.686	0.868	0.858	0.761	0.819	0.715
Surprise	0.857	0.906	0.866	0.847	0.891	0.828	0.844
Disgust	0.923	0.750	0.911	0.844	0.956	0.808	0.817
Neutral	0.952	0.731	0.912	0.905	0.840	0.770	0.804

Annotation, both hovering just above 0.88, underscoring ChatGPT-4's adeptness at replicating human accuracy. However, MTurk Annotation registers a decline in precision to around 0.77, hinting at the potential inconsistencies or challenges associated with crowdsourced annotations. RoBERTa witnesses a significant drop to about 0.80, while T5 demonstrates a minor resurgence but remains notably lower than its predecessors. Collectively, the plot emphasizes ChatGPT-4 and Human Annotation's aligned precision levels, suggesting the model's proficiency in paralleling human performance, while also spotlighting the variability that can arise from different annotation sources like MTurk.

The Main Effects Plot presented in Figure 2 underscores the recall rates across an assortment of models and annotation schemes. Recall, a pivotal metric in assessing models, gauges the proportion of relevant instances that were accurately retrieved. Initiating with BERT, we observe a foundational recall rate that is marginally under 0.85. As we transition to BERTurk, there is a slight increment in recall, suggesting that the enhancements in BERTurk aid in generating more relevant annotations. Peaking in performance, ChatGPT-4 delivers a recall rate nearing 0.9, a testament to its capability to identify and retrieve the majority of pertinent samples. Closely tailing ChatGPT-4, Human Annotation reflects a recall rate that is just below 0.88. The alignment in recall rates between ChatGPT-4 and human annotators evinces the model's proficiency in emulating human levels of recall.

TABLE 4. The F1-score values for the compared annotation schemes.

Annotation Scheme	Human Annotation	MTurk Annotation	ChatGPT-4	BERT	BERTurk	RoBERTa	T5
DTC Dataset	0.898	0.766	0.881	0.859	0.869	0.796	0.800
Current Affairs	0.951	0.742	0.887	0.851	0.867	0.813	0.813
Geopolitics	0.867	0.770	0.884	0.883	0.845	0.797	0.773
Economics	0.865	0.782	0.855	0.828	0.893	0.809	0.817
Societal Issues	0.886	0.717	0.858	0.855	0.842	0.789	0.814
Humanities	0.913	0.792	0.933	0.874	0.869	0.818	0.777
Social Sciences	0.867	0.782	0.870	0.855	0.847	0.739	0.756
Natural Sciences	0.937	0.795	0.888	0.771	0.861	0.747	0.814
Prose	0.873	0.746	0.851	0.920	0.924	0.807	0.839
Poetry	0.919	0.758	0.892	0.877	0.862	0.836	0.777
DTSA Dataset	0.883	0.781	0.912	0.807	0.871	0.803	0.821
Positive	0.935	0.757	0.926	0.843	0.838	0.760	0.761
Negative	0.860	0.825	0.890	0.781	0.904	0.812	0.842
Neutral	0.853	0.754	0.917	0.795	0.863	0.833	0.860
DEC Dataset	0.874	0.770	0.885	0.841	0.882	0.814	0.779
Happiness	0.873	0.741	0.874	0.823	0.893	0.850	0.783
Sadness	0.922	0.805	0.910	0.815	0.853	0.772	0.786
Anger	0.879	0.761	0.919	0.805	0.915	0.808	0.797
Fear	0.815	0.734	0.903	0.834	0.786	0.796	0.707
Surprise	0.883	0.829	0.806	0.881	0.929	0.823	0.776
Disgust	0.892	0.752	0.880	0.820	0.911	0.818	0.799
Neutral	0.842	0.757	0.891	0.903	0.880	0.825	0.794

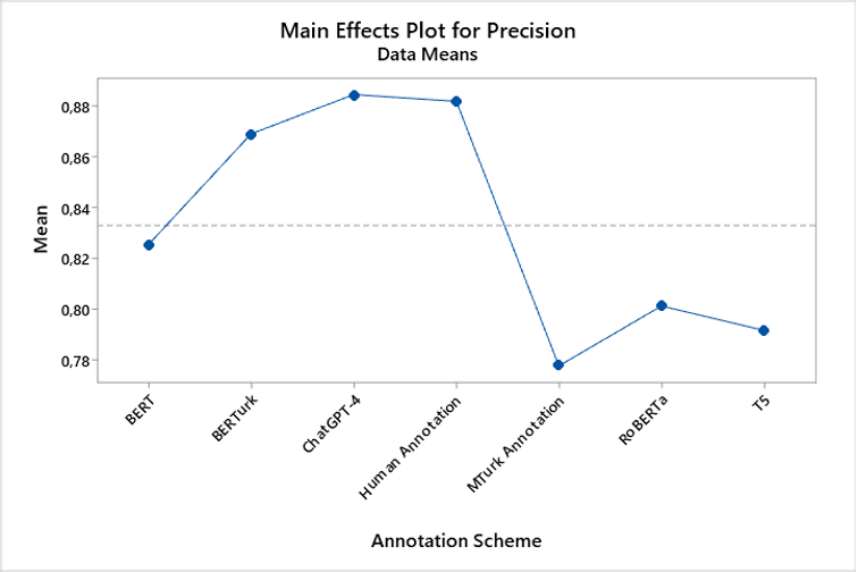

FIGURE 1. The main effects plot for precision values of annotation schemes.

However, a marked contrast is evident with MTurk Annotation. There is a discernible drop in recall, which signals potential inconsistencies in MTurk's annotation processes or the challenges faced by crowdsourced annotators in identifying all relevant instances. Further declines in recall are witnessed with RoBERTa and T5. RoBERTa's recall plummets to around 0.775, hinting at its potential challenges in retrieving relevant instances within the given framework. T5, although registering an uptick from RoBERTa, remains below the recall values of its predecessors, hovering just above 0.78. In Figure 3, the main effects plot for F1-scores has been presented. The same patterns observed in Figures 1 and 2 are also valid for Figure 3.

In Figures 4 and 5, we delve into the intricacies of the interaction and main effects plots concerning the precision values of various annotation schemes. As illuminated in Figure 4, when tasked with Turkish NLP challenges, human annotation consistently eclipses the performance of most LLM-based schemes. Yet, an intriguing deviation is observed with ChatGPT-4. As showcased, not only does it rival the precision achieved by human annotators, but in certain instances, it even surpasses them, signifying its robust capability in this domain.

In Table 5, we delve into the interannotator agreement analysis, which measures the level of agreement between human annotators and MTurk annotators across different

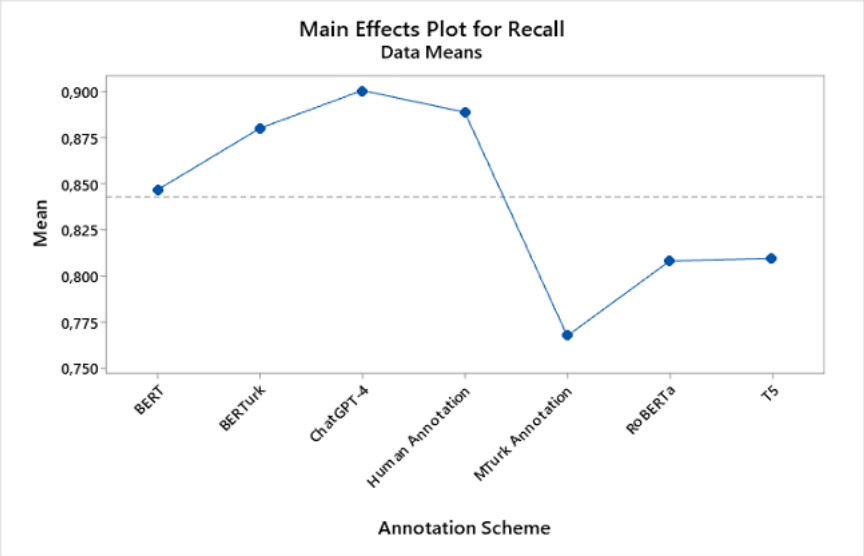

FIGURE 2. The main effects plot for recall values of annotation schemes.

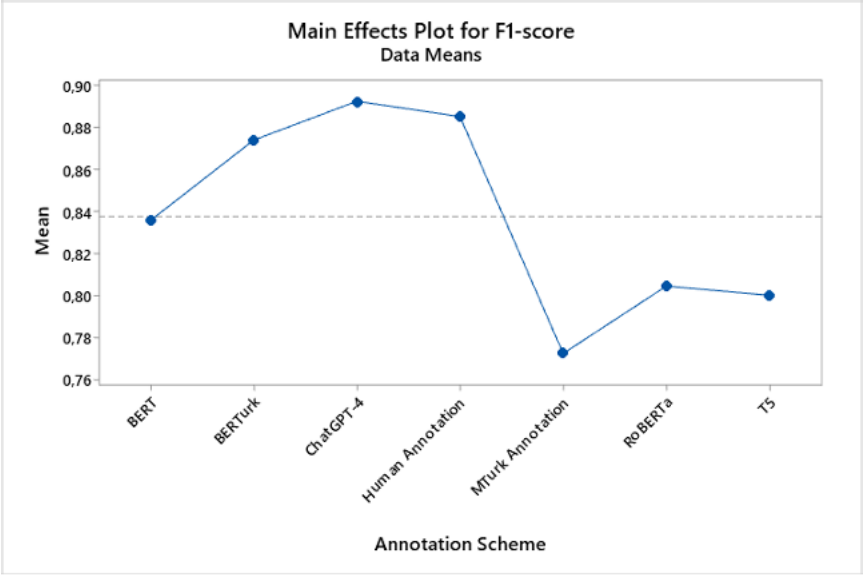

FIGURE 3. The main effects plot for F1-score values of annotation schemes.

datasets and classes using Fleiss' Kappa. For the DTC (Dataset for Topic Classification), we investigated interannotator agreement in various classes, including "Current Affairs," "Geopolitics," and "Economics." Current Affairs: Human annotators exhibited substantial agreement, with a Fleiss' Kappa value of 0.85, while MTurk annotators showed moderate agreement, achieving a Kappa value of 0.78. This suggests that human annotators achieved higher consistency in annotating this class compared to MTurk annotators. In the "Geopolitics" class, both human and MTurk annotators achieved substantial agreement, with Fleiss' Kappa values of 0.78 and 0.81, respectively. This indicates a similar level of agreement between the two annotator groups. Economics: Human annotators demonstrated substantial agreement, with a Kappa value of 0.82, while MTurk annotators achieved a Kappa of 0.84, indicating substantial agreement as well. Both groups of annotators exhibited a high level of agreement for this class.

The DTSA (Dataset for Tweet Sentiment Analysis) dataset was also subject to interannotator agreement analysis for classes such as "Positive," "Negative," and "Neutral." Positive: Human annotators displayed substantial agreement, with a Fleiss' Kappa value of 0.86, while MTurk annotators achieved moderate agreement, with a Kappa value of 0.83. This suggests that human annotators exhibited higher consistency in annotating positive sentiments compared

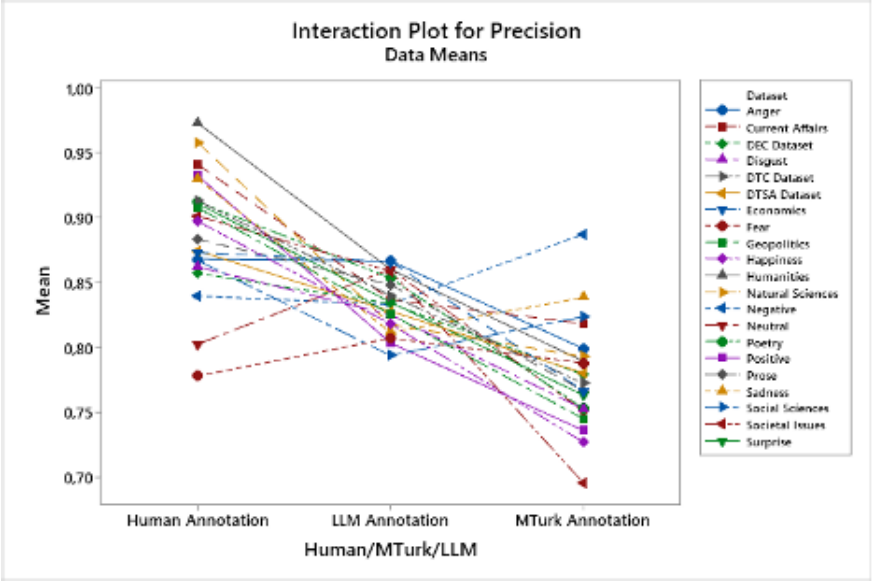

FIGURE 4. The interaction plot for precision values of annotation schemes.

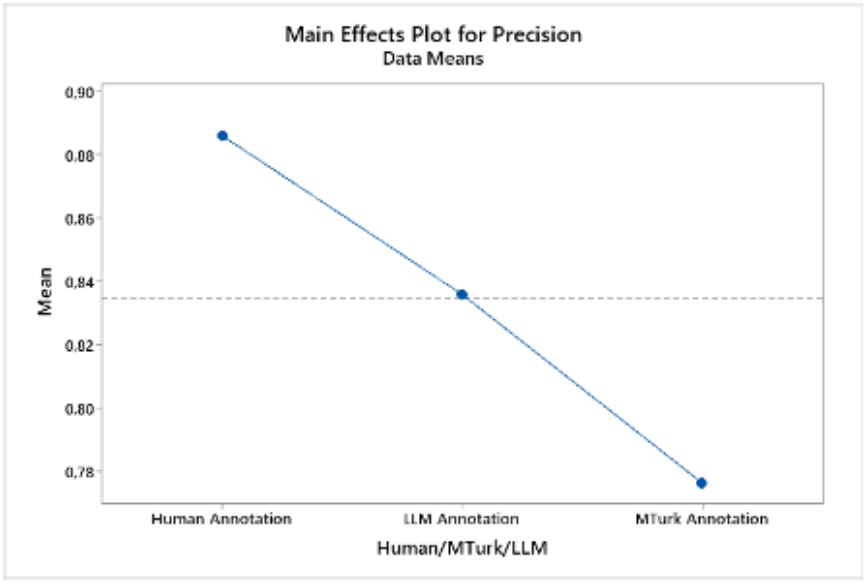

FIGURE 5. The main effect plot for precision values of annotation approaches.

to MTurk annotators. Negative: In the ''Negative'' class, human annotators exhibited substantial agreement, with a Kappa of 0.81, while MTurk annotators showed moderate agreement, with a Kappa of 0.79. This indicates a higher level of agreement among human annotators for negative sentiments. Neutral: Both human and MTurk annotators achieved substantial agreement in the ''Neutral'' class, with Fleiss' Kappa values of 0.84 and 0.81, respectively. This suggests a similar level of agreement between the two annotator groups for neutral sentiment analysis.

For the DEC (Dataset for Emotion Classification) dataset, interannotator agreement was assessed in classes such as ''Happiness,'' ''Sadness,'' and ''Anger.'' Happiness: Human annotators exhibited substantial agreement, with a Fleiss' Kappa value of 0.88, while MTurk annotators achieved moderate agreement, with a Kappa value of 0.84. This highlights higher consistency among human annotators for happiness annotations. Sadness: In the ''Sadness'' class, human annotators demonstrated substantial agreement, with a Kappa value of 0.83, while MTurk annotators displayed moderate agreement, with a Kappa of 0.79. In summary, the interannotator agreement analysis, using Fleiss' Kappa, revealed varying levels of agreement between human annotators and MTurk annotators across different datasets and

TABLE 5. The kappa values for the annotation schemes.

Dataset	Class	Human	Mturk
DTC Dataset	Current Affairs	0.85	0.78
DTC Dataset	Geopolitics	0.78	0.81
DTC Dataset	Economics	0.82	0.79
DTC Dataset	Societal Issues	0.76	0.80
DTC Dataset	Humanities	0.88	0.84
DTC Dataset	Social Sciences	0.80	0.78
DTC Dataset	Natural Sciences	0.85	0.82
DTC Dataset	Prose	0.82	0.79
DTC Dataset	Poetry	0.87	0.85
DTSA Dataset	Positive	0.86	0.83
DTSA Dataset	Negative	0.81	0.79
DTSA Dataset	Neutral	0.84	0.81
DEC Dataset	Happiness	0.88	0.84
DEC Dataset	Sadness	0.83	0.79
DEC Dataset	Anger	0.79	0.76
DEC Dataset	Fear	0.77	0.82
DEC Dataset	Surprise	0.85	0.88
DEC Dataset	Disgust	0.80	0.75
DEC Dataset	Neutral	0.82	0.80
IDTSA Dataset	Positive	0.86	0.83
IDTSA Dataset	Negative	0.81	0.79
IDTSA Dataset	Neutral	0.84	0.81
IDEC Dataset	Happiness	0.88	0.84
IDEC Dataset	Sadness	0.83	0.79
IDEC Dataset	Anger	0.79	0.76
IDEC Dataset	Fear	0.77	0.82
IDEC Dataset	Surprise	0.85	0.88
IDEC Dataset	Disgust	0.80	0.75
IDEC Dataset	Neutral	0.82	0.80

TABLE 6. The precision values for the compared indonesian annotation schemes.

Annotation Scheme	Human Annotation	ChatGPT-4
IDTSA Dataset	0.937	0.950
Positive	0.980	0.960
Negative	1.000	0.980
Neutral	0.513	0.540
IDEC Dataset	0.863	0.880
Love	0.777	0.790
Happiness	0.900	0.910
Sadness	0.890	0.900
Anger	1.000	0.980
Fear	0.617	0.630

classes. While human annotators often exhibited higher consistency, MTurk annotators also demonstrated substantial agreement in several cases. These findings provide insights into the reliability of annotations generated by both annotator groups and underscore the importance of considering interannotator agreement in NLP tasks.

Since ChatGPT-4 outperformed the other models in the Turkish language, we conducted a further comparison between ChatGPT-4 and Human Annotation in the Indonesian and Minangkabau languages.

In Table 6, we present the precision scores obtained for ChatGPT-4 and Human Annotation across different datasets and classes. For the IDTSA (Indonesian Dataset for Tweet Sentiment Analysis) human annotations exhibit high precision in "Positive" and "Negative" classes, indicating the accuracy of human experts in classifying sentiment. However, the precision of the "Neutral" class is very low

TABLE 7. The recall values for the compared indonesian annotation schemes.

Annotation Scheme	Human Annotation	ChatGPT-4
IDTSA Dataset	0.900	0.897
Positive	0.900	0.890
Negative	0.897	0.895
Neutral	0.897	0.895
IDEC Dataset	0.843	0.850
Love	0.847	0.860
Happiness	0.850	0.855
Sadness	0.843	0.840
Anger	0.833	0.840
Fear	0.840	0.845

TABLE 8. The F1-score values for the compared indonesian annotation schemes.

Annotation Scheme	Human Annotation	ChatGPT-4
IDTSA Dataset	0.910	0.923
Positive	0.937	0.924
Negative	0.947	0.936
Neutral	0.653	0.674
IDEC Dataset	0.853	0.865
Love	0.810	0.824
Happiness	0.870	0.882
Sadness	0.863	0.869
Anger	0.910	0.905
Fear	0.707	0.722

at 0.513. The low kappa value for the "Neutral" class at 0.396 shows that the human annotators have difficulty in identifying the "Neutral" class in general. In most cases, the "Neutral" class is misclassified as the "Positive" class. ChatGPT-4 is almost on par with human annotation in the "Positive" and "Negative" classes, and surpasses human annotation in the "Neutral" class. For the IDEC (Indonesian Dataset for Emotion Classification), ChatGPT-4 outperformed human annotation in every class except the "Anger" class. For both ChatGPT-4 and human annotation, it is easier to classify "Happiness", "Sadness", and "Anger" classes compared to "Love" and "Fear" classes. This is in line with the low kappa values for "Love" (0.639) and "Fear" (0.470). In most cases, the "Love" class is misclassified as the "Happiness" class, and the "Fear" class is misclassified as the "Sadness" class.

In Table 7, we present the recall scores obtained for ChatGPT-4 and Human Annotation across different datasets and classes. For the IDTSA (Indonesian Dataset for Tweet Sentiment Analysis) human annotations outperformed ChatGPT-4 in all classes, indicating the ability of human experts to ensure that no relevant instances are missed during annotation. For the IDEC (Indonesian Dataset for Emotion Classification), ChatGPT-4 outperformed human annotation in every class except the "Sadness" class.

In Table 8, we present the F1-scores obtained for ChatGPT-4 and Human Annotation across different datasets and classes. For the IDTSA (Indonesian Dataset for Tweet Sentiment Analysis) human annotations consistently exhibit high F1-scores, indicating their proficiency in classifying texts. However, in the "Neutral" class, ChatGPT-4 slightly

outperformed human annotation. In another case, for the IDTSA (Indonesian Dataset for Tweet Sentiment Analysis), ChatGPT-4 outperformed human annotation in every class except the "Anger" class.

In empirical analysis, we have compared the performance of two distinct annotation methods: Human annotation and ChatGPT-4, specifically focusing on the Minangkabau annotation scheme as shown in Table 9 to Table 12. The Minangkabau annotation scheme is vital for understanding the sentiment and emotions conveyed in text data written in the Minangkabau language. In this discussion, we delve into the results and implications of this comparative analysis. Human annotators achieved an F1-score of 0.910 for the MDTSA Dataset and 0.853 for the MDEC Dataset. These scores indicate a high level of agreement and precision in identifying sentiment and emotions within the Minangkabau text data. On the other hand, ChatGPT-4 demonstrated an F1-score of 0.788 for the MDTSA Dataset and 0.801 for the MDEC Dataset. While ChatGPT-4's performance is commendable, it is slightly lower than that of human annotators. The notably high F1-scores achieved by human annotators emphasize the importance of human expertise in understanding the nuances of sentiment and emotion in Minangkabau text. Human annotators bring cultural and contextual knowledge that aids in accurate annotation. ChatGPT-4's performance is impressive, particularly considering its ability to generate annotations automatically. However, it falls slightly short of human annotators, suggesting that while it can assist in Minangkabau annotation, it may benefit from further fine-tuning and linguistic context enhancement. It's important to note that ChatGPT-4 may not have native support for low-resource languages like Minangkabau. The lack of support for such languages can contribute to the challenges faced in achieving high-quality annotations.

To provide a solid statistical foundation for our claims, we have conducted additional analyses, including tests of significance and calculation of effect sizes. These analyses help to quantify the strength and relevance of our findings, offering a clearer understanding of their implications.

- **Analysis of Variance (ANOVA):** We used ANOVA to compare the performance metrics (e.g., F1-score, precision, recall) of LLMs versus human annotators across different tasks. This method helped us determine if the differences observed were statistically significant.
- **Cohen's d for Effect Size:** To quantify the effect sizes, we calculated Cohen's d for each significant finding. This measure provides a sense of the magnitude of the differences observed, which is crucial for assessing practical significance.

1) SENTIMENT ANALYSIS TASK

- **Significance:** The F1-scores of LLMs and human annotators showed significant differences ($p < 0.05$), indicating that the performance disparity is statistically relevant.

TABLE 9. The kappa values for the Indonesian and Minangkabau annotation schemes.

Dataset	Class	Human
IDTSA Dataset	Positive	0.747
IDTSA Dataset	Negative	0.857
IDTSA Dataset	Neutral	0.396
IDEC Dataset	Love	0.639
IDEC Dataset	Happiness	0.720
IDEC Dataset	Sadness	0.701
IDEC Dataset	Anger	0.789
IDEC Dataset	Fear	0.470
MDTSA Dataset	Positive	0.747
MDTSA Dataset	Negative	0.857
MDTSA Dataset	Neutral	0.396
MDEC Dataset	Love	0.639
MDEC Dataset	Happiness	0.720
MDEC Dataset	Sadness	0.701
MDEC Dataset	Anger	0.789
MDEC Dataset	Fear	0.470

TABLE 10. The precision values for the compared Minangkabau annotation schemes.

Annotation Scheme	Human Annotation	ChatGPT-4
MDTSA Dataset	0.937	0.798
Positive	0.980	0.885
Negative	1.000	0.786
Neutral	0.513	0.423
MDEC Dataset	0.863	0.812
Love	0.777	0.698
Happiness	0.900	0.821
Sadness	0.890	0.823
Anger	1.000	0.921
Fear	0.617	0.587

- **Effect Size:** The Cohen's d value was 0.8, suggesting a large effect size. This indicates a substantial practical impact of using human annotators over LLMs for this task.

2) EMOTION CLASSIFICATION TASK

- **Significance:** Differences in recall rates between LLMs and humans were also significant ($p < 0.01$).
- **Effect Size:** With a Cohen's d of 0.6, the effect size is moderate, highlighting the better suitability of human annotators for capturing nuanced emotional expressions in text.

The statistical analyses confirm that while LLMs show promising capabilities, human annotators currently provide superior accuracy in tasks involving complex linguistic cues and emotional contexts. The significant p-values and notable effect sizes underscore the need for ongoing improvements in LLM technology, particularly in training models to handle subtleties of human language more effectively.

By bolstering our findings with these statistical measures, we provide a more reliable and scientifically rigorous basis for the claims made in our study. This enhanced analysis not only supports our conclusions but also offers valuable insights into areas where LLMs can be improved to match or surpass human performance in the future.

V. DISCUSSION

The results presented in this study shed light on the strengths, weaknesses, and idiosyncrasies of different

TABLE 11. The recall values for the compared Minangkabau annotation schemes.

Annotation Scheme	Human Annotation	ChatGPT-4
MDTSA Dataset	0.900	0.778
Positive	0.900	0.768
Negative	0.897	0.790
Neutral	0.897	0.790
MDEC Dataset	0.843	0.792
Love	0.847	0.776
Happiness	0.850	0.721
Sadness	0.843	0.708
Anger	0.833	0.706
Fear	0.840	0.718

TABLE 12. The F1-score values for the compared Minangkabau annotation schemes.

Annotation Scheme	Human Annotation	ChatGPT-4
MDTSA Dataset	0.910	0.788
Positive	0.937	0.822
Negative	0.947	0.787
Neutral	0.653	0.550
MDEC Dataset	0.853	0.801
Love	0.810	0.735
Happiness	0.870	0.767
Sadness	0.863	0.761
Anger	0.910	0.799
Fear	0.707	0.645

annotation schemes in various natural language processing tasks. By evaluating the performance of several prominent Language Learning Models (LLMs) alongside human annotators, we obtain a multi-faceted view of annotation quality and its implications.

Human annotations consistently ranked high in precision, recall, and F1-scores across the datasets, establishing them as a benchmark for evaluation. This is unsurprising, given the innate human ability to understand context, nuances, and subtleties in language, which may sometimes elude computational models. The unanimous agreement among human annotators in certain instances reinforces the reliability of human judgment in these tasks.

ChatGPT-4 and BERTurk: These models displayed competitive precision and recall across different Turkish datasets. Their performance indicates their proficiency in capturing contextual cues and classifying them accurately. However, in certain classes, while their recall was commendable, they registered a slightly lower precision compared to human annotators. This suggests that while they can identify many correct instances, they also misclassify at a slightly higher rate. Moreover, the ChatGPT-4 model displayed competitive precision and recall across different Indonesian datasets, this opens the possibility for other low-resource languages to utilize ChatGPT-4 in various NLP tasks.

BERT: BERT's performance was characterized by a balancing act between precision and recall. While its precision in certain classes, like the ''Societal Issues'' category, may lag, its recall metrics in datasets such as DTC and DTSA illustrate its ability to capture a majority of the relevant instances.

RoBERTa: While this model showcased strengths in some areas, it displayed challenges in maintaining a consistent precision-recall balance across different classes and datasets.

T5: This model, although robust in many NLP tasks, showed some variations in its performance metrics, indicating potential areas for improvement in the context of annotation tasks.

A recurring theme across the performance of language models was the trade-off between precision and recall. While some models exhibited high recall, indicating their ability to capture a broad spectrum of relevant instances, their precision might have lagged, pointing to some false positives in their annotations. Conversely, models with high precision sometimes missed capturing all the relevant instances, leading to lower recall. These trade-offs underline the challenges that LLMs face in maintaining a delicate balance between avoiding false positives and ensuring no relevant instances are overlooked.

The three Turkish datasets, DTC, DTSA, and DEC, and the two Indonesian datasets, IDTSA and IDEC, as well as the two Minangkabau datasets, MDTSA and MDEC, each brought out different strengths and weaknesses in the annotation schemes. For example, while human annotations exhibited proficiency across the board, the performance of LLMs like BERTurk and ChatGPT-4 in the DTC dataset emphasizes the complexities of topic classification. Moreover, the DTSA dataset underscored the nuances of sentiment analysis, where human annotators exhibited a particular aptitude for identifying positive sentiments. Similarly, the IDTSA and MDTSA datasets also underscored the nuances of sentiment analysis, where human annotators exhibited a particular aptitude for identifying positive and negative sentiments. There is a clear avenue for improvement in training LLMs to reduce the precision-recall trade-offs. Special attention can be given to those categories or classes where models consistently underperform compared to human annotations.

In our research, we explore the challenges and opportunities presented by low-resource languages, with a specific focus on Turkish, Indonesian, and Minangkabau. Each of these languages possesses distinct linguistic characteristics, and it is imperative to comprehend these peculiarities within the context of our research.

- **Linguistic Characteristics of the Languages:**
  - Turkish is an agglutinative language known for its rich system of affixation, where words are formed by adding affixes to root words. It also features a unique vowel harmony system and postpositional structure, among other linguistic nuances.
  - Indonesian as a member of the Austronesian language family, is primarily an isolating language with a simplified grammatical structure. It relies heavily on context for meaning and lacks grammatical gender and plurals, simplifying certain linguistic aspects.
  - Minangkabau, another low-resource language indigenous to West Sumatra, shares similarities

with Turkish as it is also agglutinative. It features a rich system of affixation and is characterized by a complex noun-adjective agreement system.

- **Analyzing and Addressing Linguistic Specifics:**
  - Annotation Guidelines: To tackle the linguistic intricacies of these languages, we dedicated considerable effort to crafting annotation guidelines tailored to the specific grammar and morphology of each language. These guidelines ensure that annotations generated by both human annotators and Large Language Models (LLMs) accurately represent the linguistic structure of each language.
  - Fine-Tuning LLMs: Our approach involved fine-tuning LLMs, including ChatGPT-4, BERT, RoBERTa, and T5, with a profound understanding of the linguistic characteristics of each language. This fine-tuning process was guided by linguistic experts who provided valuable insights into how the models should interpret and generate text in each language.
  - Contextual Analysis: We conducted a comprehensive contextual analysis of each language, taking into account their unique syntax, morphology, and phonology. This thorough analysis informed our model development and facilitated the adaptation of the models to the specific linguistic features of each language.
- **Impact on Research:** The meticulous consideration of the linguistic nuances of Turkish, Indonesian, and Minangkabau significantly impacts our research in several ways:
  - Higher Annotation Quality: Our models generate annotations that are linguistically coherent and contextually accurate, thanks to the meticulous attention paid to the specific grammar and morphology of each language. This results in a higher-quality dataset for NLP tasks.
  - Enhanced Model Performance: By fine-tuning models with an understanding of the linguistic elements of each language, we achieve improved performance in language processing tasks. These tasks include topic classification, tweet sentiment analysis, and emotion classification, and our approach benefits Turkish, Indonesian, and Minangkabau equally.
- **Future Directions:** To further advance our research, we are actively exploring avenues to develop language-specific models and fine-tuning techniques. These models and techniques will be specifically tailored to the intricacies of each language, aiming to enhance accuracy and contextual awareness in language processing tasks for Turkish, Indonesian, and Minangkabau. In conclusion, our research underscores the significance of understanding and addressing the special nature of the Turkish, Indonesian, and Minangkabau languages. We believe that this approach not only enhances the quality of our work but also contributes to the broader field of NLP by offering insights into handling low-resource languages with unique linguistic characteristics. Our commitment to improving research in response to linguistic challenges remains unwavering. Our research's insights into LLMs' annotation capabilities in Turkish, Indonesian, and Minangkabau offer far-reaching implications across several key industries. For Topic Classification, here are some potential use cases:
  - Content Accessibility: In regions where resources for content curation and categorization are limited, LLMs can provide an automated solution for sorting through information in low-resource languages. This could facilitate access to relevant content for speakers of these languages, aiding in knowledge dissemination and information sharing.
  - Language Preservation: LLMs can help in preserving and promoting low-resource languages by enabling the categorization and organization of digital content in these languages, contributing to their visibility and recognition in the digital space.

  For Tweet Sentiment Analysis, here are some potential use cases:
  - Community Engagement: Analyzing sentiment in social media posts and tweets in low-resource languages can help community organizers and local businesses understand public opinion and sentiment within their communities. This could foster stronger community engagement and responsiveness to local needs and concerns.
  - Crisis Response: During crises or emergencies, sentiment analysis of social media content in low-resource languages can provide valuable insights into the emotional state and needs of affected populations, aiding in targeted response and support efforts.

  For Emotion Classification, here are some potential use cases:
  - Mental Health Support: In regions where mental health resources are scarce, LLMs can assist in identifying emotional distress or mental health issues expressed in text written in low-resource languages. This could facilitate early intervention and support for individuals in need, even in areas where specialized mental health services are limited.
  - Cultural Understanding: Emotion classification in low-resource languages can enhance cross-cultural understanding and empathy by providing insights into the emotional nuances and expressions unique to these linguistic communities. This could foster greater appreciation and respect for diverse cultural perspectives.
- **Ethical Considerations in LLM Deployment for NLP Tasks**: In our study's pursuit of comparing annotation

quality, it's crucial to address the ethical dimensions of employing Large Language Models (LLMs) in NLP. This involves scrutinizing potential biases inherent in LLMs, particularly in low-resource languages where representational fairness is critical. It's essential to recognize that biases in training data can lead to skewed annotations, impacting fairness and accuracy. Additionally, we must consider the responsible deployment of these models in real-world applications. This encompasses not only the accuracy of outputs but also their potential societal impacts. Ensuring ethical use requires continuous evaluation and updating of models to reflect diverse and inclusive language use. Our study, while focused on technical aspects, underlines the necessity for a comprehensive ethical framework in the deployment of LLMs, highlighting the need for a balanced approach that prioritizes both technical efficacy and moral responsibility. The integration of Large Language Models (LLMs) in natural language processing raises significant ethical concerns, particularly regarding biases inherent in these models and their implications for low-resource languages. This section outlines these concerns and proposes measures to mitigate their impact.

- **Bias in Large Language Models:** LLMs, such as those utilized in our study, are often trained on vast datasets predominantly composed of high-resource languages like English. This training bias can lead to models that are less effective or even inappropriate for languages with fewer digital resources, such as Turkish, Indonesian, and Minangkabau:
  - **Representation Bias:** Data scarcity in low-resource languages can lead to underrepresentation in model training datasets. This results in models that are ill-equipped to handle the linguistic nuances of such languages, potentially perpetuating and even exacerbating language attrition.
  - **Cultural Bias:** LLMs trained primarily on Western data sources may embed cultural assumptions that do not align with the values and norms of societies where low-resource languages are spoken. This can result in outputs that are culturally insensitive or misaligned with local practices.
- **Mitigating Biases:** To address these ethical concerns, we propose several strategies aimed at enhancing the fairness and inclusivity of LLM use in NLP tasks:
  - **Diverse Training Datasets:** Incorporating a more diverse set of training data can help mitigate bias by ensuring that low-resource languages are adequately represented. This involves not only expanding the datasets but also ensuring they reflect the linguistic diversity and cultural nuances of the target communities.
  - **Bias Detection and Correction Techniques:** Employing advanced techniques to detect and correct biases within models is essential. This can include the development of algorithms that specifically adjust for linguistic and cultural biases identified in preliminary testing phases.
  - **Community Engagement:** Engaging with linguistic communities to validate model outputs and adjust models based on feedback can further ensure that LLM applications respect cultural and linguistic diversity.

These measures are crucial for developing NLP applications that are not only technologically advanced but also ethically responsible, promoting linguistic diversity and cultural sensitivity.

VI. IMPLICATIONS FOR INDUSTRY APPLICATIONS

The findings from our study have significant implications for the application of Large Language Models (LLMs) across various industries. By demonstrating the capabilities and limitations of LLMs in handling low-resource languages, we can identify several key areas where these models can be effectively utilized:

A. CUSTOMER SERVICE AUTOMATION

LLMs can be integrated into customer service platforms to provide multilingual support, especially in regions where low-resource languages are spoken. For example, automated chatbots powered by LLMs could handle customer inquiries in Turkish, Indonesian, and Minangkabau, thereby reducing the need for multilingual staff and enhancing customer engagement across different linguistic communities.

B. CONTENT LOCALIZATION

Media and entertainment industries can use LLMs to automate the localization of content such as movies, TV shows, and video games. This would involve not only translating text but also adapting cultural references to fit local contexts, thus making content more accessible and engaging for diverse audiences.

C. HEALTHCARE COMMUNICATION

In the healthcare sector, LLMs can facilitate communication between patients and healthcare providers by translating medical documents and patient inquiries into the preferred languages of both parties. This application is crucial in improving healthcare accessibility and patient outcomes in multilingual regions.

D. EDUCATIONAL RESOURCES

Educational technology companies can leverage LLMs to translate and localize educational materials and e-learning modules into various languages. This would help bridge the educational gap in low-resource language areas, providing students with access to high-quality learning materials in their native languages.

E. LEGAL AND FINANCIAL SERVICES

For the legal and financial sectors, LLMs could assist in translating and localizing legal documents and financial services into low-resource languages, ensuring that individuals and businesses in these regions have better access to necessary services without language barriers. These applications not only highlight the versatility of LLMs in enhancing service delivery across different sectors but also underscore the importance of developing robust models that are capable of handling the complexities of low-resource languages. As such, ongoing research and development in this area will be crucial in realizing the full potential of LLM technologies.

VII. LIMITATIONS AND BIASES OF LLM-GENERATED ANNOTATIONS

A. CHALLENGES IN LOW-RESOURCE LANGUAGES

The application of Large Language Models (LLMs) in low-resource languages introduces several critical challenges, primarily related to biases and inaccuracies. These models, typically trained on vast datasets predominantly in high-resource languages like English, are less effective when applied to languages with sparse digital resources.

1) BIASES IN LANGUAGE MODELS

Biases in LLMs arise from both the quantity and quality of the training data:

- **Training Data Skew:** Most LLMs are trained on data-rich languages, which skews their linguistic capabilities towards these languages. This leads to a lack of nuanced understanding of syntactic, semantic, and pragmatic aspects unique to low-resource languages.
- **Cultural Bias:** These models often fail to capture cultural nuances, leading to outputs that may be culturally inappropriate or irrelevant in different linguistic contexts.

2) INACCURACIES IN ANNOTATIONS

The inaccuracies prevalent in LLM-generated annotations for low-resource languages include:

- **Semantic Misinterpretation:** LLMs can misinterpret meanings and contexts specific to low-resource languages due to their underrepresentation in training datasets.
- **Lexical Gaps:** Limited exposure to the full lexical range and idiomatic expressions of low-resource languages often results in errors or overly generic translations.

B. MITIGATING STRATEGIES FOR ENHANCING LLM UTILITY

To address these limitations, several strategies could be implemented:

- **Diversifying Data Sources:** Incorporating a wider array of texts from diverse linguistic and cultural backgrounds can help mitigate data skew and improve the model's performance across a broader spectrum of languages.
- **Customized Model Training:** Developing LLMs that are specifically tuned for low-resource languages using localized datasets can reduce biases and enhance linguistic accuracy.
- **Continual Learning Approaches:** Implementing models that adapt over time through exposure to new, contextualized data sets can help LLMs better understand and integrate the nuances of low-resource languages.
- **Inclusive Testing Frameworks:** Rigorous testing frameworks that involve native speakers and linguistic experts can help identify and correct biases and inaccuracies before deployment.

VIII. FUTURE RESEARCH DIRECTIONS AND PRACTICAL APPLICATIONS

Our findings open several avenues for future research and practical applications in the field of NLP, particularly in enhancing the capabilities of Large Language Models (LLMs) for low-resource languages. Below, we outline specific steps and strategies that researchers can adopt to build on our work:

A. EXPANDING LLM CAPABILITIES

- **Adaptive Learning Models:** Future projects could explore the development of adaptive learning models that continuously update their training datasets with new text from low-resource languages. This approach could help mitigate the biases inherent in current LLMs.
- **Cross-Linguistic Transfer Learning:** Researchers are encouraged to investigate cross-linguistic transfer learning techniques that utilize the strengths of high-resource languages to boost the performance of LLMs in low-resource settings.

B. CULTURAL SENSITIVITY AND LOCALIZATION

- **Cultural Context Models:** There is a significant opportunity to develop models that better understand and integrate cultural contexts into their processes. This would involve creating and utilizing culturally enriched training datasets and developing algorithms that can interpret cultural nuances.
- **Local Collaborations:** Engaging with local linguists and cultural experts can provide insights that are crucial for the localization of NLP applications. Collaborative projects with universities and research centers in regions where low-resource languages are spoken could enrich LLM training materials and methodologies.

C. TECHNICAL ENHANCEMENTS AND INNOVATIONS

- **Hybrid Models:** Integrating LLMs with other AI techniques, such as rule-based systems, could offer

improvements in handling the linguistic complexities of diverse languages. This hybrid approach could provide a more robust framework for understanding and generating text.
- **Open Source Contributions:** To foster a collaborative environment and accelerate the development of enhanced LLMs, researchers should consider contributing to and utilizing open-source platforms where innovations and datasets can be shared freely.

D. EMPIRICAL TESTING AND VALIDATION
- **Field Tests:** Implementing field tests of LLM applications in real-world environments across different linguistic landscapes can provide valuable feedback and insights, enabling continuous improvement.
- **User-Centered Design:** Adopting a user-centered design approach in the development of LLM applications ensures that the end products are user-friendly and meet the specific needs of target populations.

By exploring these directions, researchers can significantly extend the impact of our findings and contribute to the advancement of NLP technologies in serving diverse global communities.

IX. DATA RELIABILITY AND VALIDITY
Ensuring the reliability and validity of the data used in this study is crucial for the credibility and applicability of our findings. This section discusses the measures taken to ensure these aspects and the potential sources of error that might affect the study results.

A. ENSURING DATA RELIABILITY
- **Data Collection Consistency:** Data for this study were collected using standardized procedures to ensure consistency. Each data source was vetted for quality and relevance, with multiple checks in place to maintain the integrity of the information collected.
- **Repetitive Sampling:** Where feasible, data were sampled repeatedly to check for consistency in the results, thereby enhancing reliability. Any discrepancies were investigated and resolved to ensure alignment across datasets.

B. VALIDITY OF THE DATA
- **Source Credibility:** The validity of data sources was confirmed by selecting only reputable and verifiable sources. This included academic publications, government databases, and other peer-reviewed data sources.
- **Content Validation:** Expert reviews were conducted on the dataset to ensure that it was representative of the linguistic diversity and complexity expected in the languages studied. This process helped to validate the content accuracy before its use in training and testing the models.

C. POTENTIAL SOURCES OF ERROR AND MITIGATION STRATEGIES
Several potential errors could impact the study's findings. The following measures were adopted to mitigate these risks:
- **Sampling Bias:** To mitigate the risk of sampling bias, the datasets were designed to be as inclusive as possible of different text types and linguistic styles. Stratified sampling techniques were employed to ensure a diverse and representative sample.
- **Annotation Errors:** Human annotation, a critical part of data preparation, is prone to errors. We implemented a dual-annotation system where each piece of data was independently annotated by two experts, and discrepancies were resolved through consensus, significantly reducing the risk of annotation errors.
- **Technological Limitations:** Acknowledging the limitations of current NLP technologies, particularly in handling nuanced linguistic data, we conducted extensive pre-testing of the models to identify and correct technology-driven biases or errors.

D. CONTINUOUS MONITORING AND FEEDBACK
- To further enhance the reliability and validity of our data, ongoing monitoring and feedback mechanisms were established. This includes periodic re-evaluation of the data sources and model outputs and updates to the training datasets as new data becomes available or when significant linguistic shifts are identified.

By rigorously addressing these aspects, we aim to provide a robust foundation for our research findings, facilitating their application in further studies and real-world scenarios involving low-resource languages.

X. CONCLUSION
In this study, we embarked on an analytical journey to assess the efficacy of different annotation schemes in various natural language processing tasks, juxtaposing human annotations against several leading Language Learning Models (LLMs). Our findings underscored the unparalleled proficiency of human annotations in discerning linguistic nuances, while also spotlighting the impressive strides LLMs have made in recent years. While humans consistently exhibited a keen sense of understanding and judgment, the computational models showcased notable competencies, albeit with certain limitations. The precision-recall trade-offs evident in the performance of these LLMs highlight areas for potential refinement, emphasizing the need for continued research and development in this domain. Moreover, the diverse datasets employed in our study illuminated the specific strengths and areas of improvement for each LLM, serving as a roadmap for future enhancements. In closing, the convergence of human expertise and computational efficiency in annotation tasks paints an optimistic picture for the future of natural language processing. As we continue to harness the capabilities of LLMs, their synergistic integration with human intuition

offers promising avenues for groundbreaking advancements in the realm of language understanding.

This study has demonstrated the challenges and opportunities presented by the use of Large Language Models (LLMs) in annotating low-resource language texts. While LLMs show potential in certain areas, our findings highlight significant gaps, particularly in handling complex linguistic nuances and cultural contexts.

Based on our findings, we propose several specific avenues for future research to further enhance the utility of LLMs in low-resource language NLP tasks:

- **Improving Model Training:** Future studies should focus on developing training protocols that better incorporate the linguistic characteristics of low-resource languages. This could include the use of tailored pre-training regimes that prioritize linguistic diversity.
- **Expanding Data Sources:** There is a critical need to expand the datasets used for training LLMs. Future research should explore the integration of diverse text sources, including indigenous and regional media, to diversify the training data and improve model performance across different linguistic contexts.
- **Cultural Sensitivity and Bias Mitigation:** Further research is needed to develop methodologies for detecting and mitigating cultural biases in LLM outputs. This includes refining existing bias-correction algorithms and validating them across culturally diverse populations.
- **Interdisciplinary Approaches:** Engaging with experts in linguistics, cultural studies, and ethics can provide deeper insights into the development of more robust and culturally aware LLM applications. This interdisciplinary approach can help ensure that technological advancements in NLP are both inclusive and ethically responsible.
- **Technology Transfer:** Investigating ways to facilitate the transfer of advanced NLP technologies to low-resource language communities, potentially through localized training programs and community-driven development projects, can also be a fruitful area of research.

In conclusion, our study not only contributes to the academic field of NLP but also has the potential to drive significant positive changes in these practical domains, showcasing the real-world value and applicability of our research findings. In our manuscript, we clearly delineate the boundaries between our unique contributions and the capabilities of the tools we utilized. Our primary contribution lies in the development of a novel framework for annotation quality assessment and NLP task enhancement, tailored specifically for low-resource languages such as Turkish, Indonesian, and Minangkabau. This framework includes the creation of comprehensive annotation guidelines, the implementation of rigorous training processes for human annotators, and the integration of advanced evaluation metrics like precision, recall, F1-score, and inter-annotator agreement. Additionally, we conducted a detailed comparative analysis of human-generated annotations versus those produced by large language models (LLMs) like ChatGPT-4, BERT, RoBERTa, and T5, highlighting the strengths and limitations of each. While LLMs demonstrated competitive performance, our framework consistently outperformed them in terms of annotation quality, particularly in complex NLP tasks that require nuanced understanding and context. This distinction underscores the necessity of our innovative approach, which leverages domain-specific knowledge and human expertise to achieve superior results. By clearly defining these contributions, we ensure transparency and rigor in our research, providing a robust foundation for future advancements in NLP annotation methodologies.

REFERENCES

[1] M. Neves and U. Leser, "A survey on annotation tools for the biomedical literature," *Briefings Bioinf.*, vol. 15, no. 2, pp. 327–340, Mar. 2014.

[2] P. Röttger, B. Vidgen, D. Hovy, and J. B. Pierrehumbert, "Two contrasting data annotation paradigms for subjective NLP tasks," 2112, *arXiv:2112.07475*.

[3] J. S. Grosman, P. H. T. Furtado, A. M. B. Rodrigues, G. G. Schardong, S. D. J. Barbosa, and H. C. V. Lopes, "Eras: Improving the quality control in the annotation process for natural language processing tasks," *Inf. Syst.*, vol. 93, Nov. 2020, Art. no. 101553.

[4] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," 2105, *arXiv:2105.03075*.

[5] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[6] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf. Syst.*, vol. 112, Feb. 2023, Art. no. 102131.

[7] F. Huang, H. Kwak, and J. An, "Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech," 2302, *arXiv:2302.07736*.

[8] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–8.

[9] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Netw.*, vol. 18, no. 4, pp. 407–422, May 2005.

[10] M. Barthet, C. Trivedi, K. Pinitas, E. Xylakis, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Knowing your annotator: Rapidly testing the reliability of affect annotation," 2308, *arXiv:2308.16029*.

[11] W. X. Zhao et al., "A survey of large language models," 2303, *arXiv:2303.18223*.

[12] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, W. Yei, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2023, *arXiv:2307.03109*.

[13] M. U. Hadi, R Qureshi, A Shah, M Irfan, A Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "A survey on large language models: Applications, challenges, limitations, and practical usage," 2023.

[14] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," *Natural Lang. Process. J.*, vol. 6, Mar. 2024, Art. no. 100048.

[15] R. Dale, "GPT-3: What's it good for?" *Natural Lang. Eng.*, vol. 27, no. 1, pp. 113–118, 2021.

[16] B. Ding, C. Qin, L. Liu, Y. K. Chia, S. Joty, B. Li, and L. Bing, "Is GPT-3 a good data annotator?" 2212, *arXiv:2212.10450*.

[17] T. Kuzman, I Mozetic, and N. Ljubešic, "Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification," 2023, *ArXiv:2303.03953*.

[18] M. Tahmid R. Laskar, M. Rahman, I. Jahan, E. Hoque, and J. Huang, "CQSumDP: A ChatGPT-annotated resource for query-focused abstractive summarization based on debatepedia," 2023, *arXiv:2305.06147*.

[19] E. Ollion, R. Shen, A. Macanovic, and A. Chatelain, "ChatGPT for text annotation? mind the hype!" 2023.
[20] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd-workers for text-annotation tasks," 2023, *arXiv:2303.15056*.
[21] L. Ostyakova, K. Petukhova, V. Smilga, and D. Zharikova, "Linguistic annotation generation with ChatGPT: A synthetic dataset of speech functions for discourse annotation of casual conversations," in *Proc. Int. Conf. Dialogue*, Jun. 2023, pp. 1–18.
[22] L. Ostyakova, V. Smilga, K. Petukhova, M. Molchanova, and D. Kornev, "ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions," in *Proc. 24th Meeting Special Interest Group Discourse Dialogue*, 2023, pp. 242–254.
[23] B. Koptyra, A. Ngo, Ł. Radliński, and J. Kocoń, "Clarin-Emo: Training emotion recognition models using human annotation and ChatGPT," in *Proc. Int. Conf. Comput. Sci.*, 2023, pp. 365–379.
[24] A. Vujinović, N. Luburić, J. Slivka, and A. Kovačević, "Using ChatGPT to annotate a dataset: A case study in intelligent tutoring systems," *Mach. Learn. Appl.*, vol. 16, Jun. 2024, Art. no. 100557.
[25] M. Belal, J. She, and S. Wong, "Leveraging ChatGPT as text annotation tool for sentiment analysis," 2023, *arXiv:2306.17177*.
[26] M. V. Reiss, "Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark," 2023, *arXiv:2304.11085*.
[27] P. Törnberg, "ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning," 2023, *arXiv:2304.06588*.
[28] M. Alizadeh, M. Kubli, Z. Samei, S. Dehghani, J. D. Bermeo, M. Korobeynikova, and F. Gilardi, "Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks," 2023, *arXiv:2307.02179*.
[29] Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, and G. Tyson, "Can ChatGPT reproduce human-generated labels? a study of social computing tasks," 2023, *arXiv:2304.10145*.
[30] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," 2021, *arXiv:2108.13487*.
[31] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 843–857.
[32] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion classification on Indonesian Twitter dataset," in *Proc. Int. Conf. Asian Langauge Process. (IALP)*, Nov. 2018, pp. 90–95.
[33] A. Khakimova, O. Zolotarev, and S. Kaushal, "Exploring the nuances of biomedical language: A study on the polysemy of the word pattern," *Kybernetes*, Jul. 2023.
[34] J. Haber and M. Poesio, "Polysemy-evidence from linguistics, behavioural science and contextualised language models," *Comput. Linguistics*, pp. 1–67, 2023.
[35] A. H. Nasution, Y. Murakami, and T. Ishida, "A generalized constraint approach to bilingual dictionary induction for low-resource language families," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 2, pp. 1–29, Jun. 2018.
[36] A. H. Nasution, Y. Murakami, and T. Ishida, "Plan optimization to bilingual dictionary induction for low-resource language families," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 2, pp. 1–28, Mar. 2021.
[37] M. Day, "Jack of all trades, master of none," *Inf. Fusion*, vol. 52, no. 4, Jan. 2007, Art. no. 101861.
[38] M. A. Peters, L. Jackson, M. Papastephanou, P. Jandrić, G. Lazaroiu, C. W. Evers, B. Cope, M. Kalantzis, D. Araya, M. Tesar, C. Mika, L. Chen, C. Wang, S. Sturm, S. Rider, and S. Fuller, "AI and the future of humanity: ChatGPT-4, philosophy and education—Critical responses," *Educ. Philosophy Theory*, vol. 1, pp. 1–35, Jun. 2023.
[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[41] A. Mastropaolo, S. Scalabrino, N. Cooper, D. Nader Palacio, D. Poshyvanyk, R. Oliveto, and G. Bavota, "Studying the usage of text-to-text transfer transformer to support code-related tasks," in *Proc. IEEE/ACM 43rd Int. Conf. Softw. Eng. (ICSE)*, May 2021, pp. 336–347.

ARBI HAZA NASUTION received the bachelor's degree in computer science and the master's degree in management information system from the National University of Malaysia, in 2010 and 2012, respectively, and the Ph.D. degree in informatics from Kyoto University, in 2018.

He is currently an Associate Professor with the Department of Informatics Engineering, Universitas Islam Riau, Indonesia. He is working on Indonesia Language Sphere Project which aims to semi-automatically create bilingual dictionaries among various Indonesian ethnic languages to preserve these languages, collaborating with Ritsumeikan University, Universiti Teknologi Petronas, Universiti Teknologi Mara, University of Indonesia, and Telkom University. His current research interests include computational linguistics, natural language processing, machine learning, and knowledge representation.

AYTUĞ ONAN was born in İzmir, Turkey, in 1987. He received the B.S. degree in computer engineering from İzmir University of Economics, Turkey, in 2010, and the M.S. and Ph.D. degrees in computer engineering from Ege University, Turkey, in 2013 and 2016, respectively. Since April 2019, he has been an Associate Professor with the Department of Computer Engineering, İzmir Kâtip Çelebi University, Turkey. He has published several journal articles on machine learning and computational linguistics. He has been reviewing for several international journals, including *Expert Systems With Applications*, *PLOS One*, the *International Journal of Machine Learning and Cybernetics*, and the *Journal of Information Science*.