Human Versus AI: A Comparative Study of Zero-Shot LLMs and Transformer Models Against Human Annotations for Arabic Sentiment Analysis

Dimah Alahmadi
Information Systems Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Abstract—**Accurate sentiment analysis in Arabic natural language processing (NLP) remains a complex task due to the language's rich morphology, syntactic variability, and diverse dialects. Traditional annotation approaches require human experts, face significant challenges related to inter-annotator agreement and dialectal understanding. Recent advances in transformer-based models and large language models (LLMs) offer new techniques to generate annotations. This paper presents a comparative evaluation of three sentiment annotation strategies applied to Saudi dialect tweets: human expert labeling, fine-tuned transformer models (specifically CAMeLBERT-DA), and zero-shot inference using GPT-4o. The selected CAMeLBERT-DA which is already trained specifically for Arabic sentiment tasks and dialects, demonstrates robust performance with fast, scalable predictions. On the other hand, the selected GPT-4o shows competitive zero-shot accuracy without fine-tuning, making it a practical solution for real-time applications. We investigate how each approach performs on two datasets, both of more than 4,000 Saudi tweets covering a wide spectrum of dialects and sentiment expressions. Our methodology involves analyzing consistency across annotations using inter-rater agreement metrics such as Cohen's Kappa, Pearson correlation, and class-specific agreement rates. The results reveal that while human annotations capture cultural and context subtleties, they suffer from inconsistency, particularly in ambiguous or dialect-specific cases. This study contributes to the growing body of work on annotation methodologies by highlighting the strengths and limitations of both human and AI-based annotators in Arabic NLP. Our findings suggest that the zero-shot use of domain-specific transformers like CAMeLBERT-DA with general-purpose LLMs such as GPT-4o have a moderate correlation compared to actual human annotators. The paper concludes with recommendations for building reliable ground truth datasets and integrating AI-assisted labeling into Arabic NLP tasks.**

Keywords*—*Large Language Model (LLMs); transformers; NLP; annotation; inter agreement; sentiment analysis, Saudi dialects

I. Introduction

The rapid advancement of large language models (LLMs) has transformed natural language processing (NLP) across multiple languages, including Arabic. Despite this progress, Arabic remains underrepresented and presents unique challenges due to its complex morphology, diglossia (the co-existence of Modern Standard Arabic and diverse dialects), and wide regional variation [1] and [2]. These linguistic characteristics make high-quality annotation for Arabic NLP tasks particularly demanding. The worldwide use and exchange of data with diverse platforms facilitate large-scale interactions between individuals. This resulted in a volume of text data where NLP plays a critical role in the future of smart applications. The advancement in NLP techniques helps to maintain and analyze such a volume of text data to extract meaningful decisions similar to humans' processing in areas like machine translation, text summarization, question answering, and sentiment analysis.

Manual annotation by expert linguists is often considered the gold standard for creating reliable datasets. However, this process is resource-intensive, time-consuming, and difficult to scale—especially when applied to informal and diverse content like social media. In recent years, researchers have turned to large language models (LLMs), such as OpenAI's GPT series for different tasks in NLP, such as sentiment analysis and machine translation [3], GPTs have a potential solution to reduce the burden of manual annotation. These models, particularly in zero-shot and few-shot settings, have shown remarkable success in performing NLP tasks without domain-specific training, raising questions about their reliability as annotation tools.

In the context of Arabic NLP, the utility of LLMs for sentiment analysis and dialect-sensitive tasks has not been comprehensively evaluated. Additionally, there is a need to benchmark the quality of their outputs against those produced by human annotators. This is particularly crucial when LLMs are used to process social media data, where language use tends to be informal, ambiguous, and culturally nuanced [4].

This study investigates the effectiveness of zero-shot GPT-4o in predicting Arabic sentiment analysis and fine-tuned transformer models (e.g., CAMeLBERT-DA) ability compared to human expert annotations. We aim to explore the validity and practical trade-offs of each annotation methodology.

The research is guided by the following questions:

- How do GPT-4o and CAMeLBERT-DA perform in zero-shot sentiment classification of Saudi dialect tweets compared to human annotation?
- How do annotation metrics differ in terms of agreement and correlation for real-world Arabic dialect social media datasets?

By addressing these questions, the study contributes to the growing body of research on efficient informed annotation practices in low-resource and linguistically diverse settings such as Arabic NLP.

A. Literature Review

Recent studies have highlighted several research gaps in this domain. Notably, there is a lack of thorough investigation of the annotation quality produced by LLMs for Arabic NLP tasks, particularly across different dialects and task types. While some work has benchmarked LLM performance against traditional models [5]. In addition, few have directly compared LLM-generated labels to human-annotated gold standards in Arabic and examined sources of error in LLM-generated annotations and transformers, as well as where they diverge most from human judgments.

B. Annotation Challenges in Arabic NLP

Arabic presents unique annotation challenges in NLP due to its morphological complexity, rich syntax, and multiple dialects that deviate significantly from Modern Standard Arabic (MSA) [6]. Arabic is a polyglossic language comprising Modern Standard Arabic (MSA) and numerous regional dialects. These dialects differ significantly in phonology, morphology, lexicon, and syntax, making unified dataset construction particularly difficult [7]. Annotators must often be familiar with both MSA and dialectal varieties to accurately label sentiment, intent, or named entities, especially when working with informal texts like tweets or online forums.

Another key challenge lies in the scarcity of large-scale, publicly available annotated datasets. Unlike English, where benchmarks like GLUE and SuperGLUE have standardized evaluation protocols, Arabic NLP still suffers from fragmented resources [8]. Many datasets are restricted to specific genres such as news propaganda, limiting their utility in social or conversational contexts on X [9].

Moreover, the presence of ambiguous structures, mixed-code usage (e.g., Arabic-English), and sarcasm prevalent in social media further complicate the annotation process. These linguistic phenomena require not only syntactic understanding but also pragmatic and cultural awareness. Thus, high inter-annotator agreement is difficult to achieve, particularly when annotations are conducted by non-experts [10].

Another major challenge is the high-quality annotated datasets in Arabic. Many Arabic NLP tasks – from hate speech detection to sentiment analysis suffer from trust labeled data, especially compared to English [11]. This is partly due to the labor-intensive nature of creating Arabic corpora: manual annotation is time-consuming, expensive, and often requires domain or dialect experts. The process can be prone to inconsistencies as well; human annotators sometimes disagree on labels for complex or ambiguous texts, introducing variability into the dataset. Arabic's writing system (e.g. omission of short vowels) and extensive morphological ambiguity can further complicate annotation, since the correct interpretation of a word or sentence may rely on context or annotator intuition. In tasks like offensive language detection, the lack of diacritics and the prevalence of informal expressions make it hard for annotators to determine intent or meaning without context [12]. Authors in study [13] have presented blockchain technology to generate a wide trust annotated dataset by crowdsourcing applications that are based on blockchain structures.

In summary, the literature indicates that Arabic NLP faces persistent annotation challenges: scarcity of labeled data, dialect and domain mismatches between annotators and content, and the inherent linguistic complexity of Arabic. These issues need careful dataset design (with clear guidelines and training for annotators) and innovative approaches to support or complement human annotators, as we explore in subsequent sections.

C. Transformer and GPT Models for Arabic NLP

In response to these limitations, large language models (LLMs) have emerged as a promising alternative to manual annotation. Transformer-based models such as BERT [14] and T5 [15] have been adapted for Arabic via pretraining on Arabic corpora.

Transformer-based models have revolutionized Arabic NLP. Initial efforts like AraBERT and Arabic-BERT demonstrated strong performance gains by pretraining on large Arabic corpora [16], [17]. Later, MARBERT and ARBERT incorporated dialectal Arabic into their training to improve generalization across informal text [6]. These models consistently outperformed multilingual baselines on downstream tasks.

Notably, CAMeLBERT-DA [18] was trained on the CAMeL Lab's extensive Arabic text collections and fine-tuned for dialectal sentiment classification, showing state-of-the-art performance on multiple downstream tasks.

Meanwhile, multilingual and instruction-tuned models like mT5 [19] and XGLM [20] have shown competence in zero-shot tasks. These models, though not Arabic-specific, benefit from broad cross-lingual transfer. However, they still underperform on dialectal Arabic unless further adapted or fine-tuned.

With the release of ChatGPT and the GPT-4 series [21], zero-shot and few-shot prompting has gained traction. These models exhibit impressive ability to generalize across tasks, offering a practical solution for rapid annotation. However, their performance in Arabic sentiment annotation is still underexplored, especially in informal settings. Recent work by Khalifa et al. [22] highlighted that GPT-4 can approximate expert annotation on formal Arabic , but may struggle with dialectal variation and ambiguous expressions.

Transformer models have not been limited to encoders. Researchers also explored generative transformers for Arabic. For instance, an AraGPT2 model (an Arabic version of GPT-2) was developed as a language generator [23]. AraGPT2 has been used alongside BERT models in hybrid systems; one study combined CAMeLBERT (encoder) with AraGPT2 (generator) in an ensemble, achieving superior accuracy on an Arabic text classification task. However, compared to the proliferation of Arabic BERT models, generative pre-trained transformers dedicated to Arabic are fewer. This is changing with the rise of massive multilingual models like GPT-3 and GPT-4, which, despite not being Arabic-specific, are trained on enough Arabic text to demonstrate impressive capabilities in the language. These GPT-series models can perform Arabic tasks in a zero-shot or few-shot manner – a fundamentally different paradigm from task-specific fine-tuning. Researchers have been keen to evaluate how well such models handle Arabic understanding and generation.

In a comprehensive evaluation across 60+ Arabic NLP benchmarks, Khondaker et al. found that ChatGPT's zero-shot accuracy was consistently lower than that of dedicated Arabic models on tasks like text classification, NER, and QA [24]. These models enable broader experimentation in Arabic NLP, but challenges remain in understanding their bias and limitations across dialects and domains. These findings suggest that while large GPT-style models are very powerful, Arabic-specific adaptation (through fine-tuning or prompting techniques) remains important for top performance. Nevertheless, the convenience of prompting an LLM without needing task-specific training is highly attractive, particularly for addressing the data scarcity issue. This has led to explorations of using GPT models as annotation tools, which we discuss in this research.

D. Hybrid Annotation Approaches

Hybrid annotation workflows are a growing area of interest. In this approach, LLMs generate initial labels that are then validated or corrected by human experts. Studies in English NLP [25] show this reduces cost and improves consistency. This framework remains under-investigated in Arabic NLP, despite its potential for scalable, culturally sensitive annotation. There is a pressing need for empirical studies that evaluate such workflows across diverse dialects, task types, and prompt strategies. One approach is to use LLMs to generate initial labels or data, which are then verified by humans. For example, Senator et al. employed ChatGPT to perform semantic role labeling (SRL) on Arabic texts. In their study, ChatGPT was prompted to assign semantic role tags to Arabic sentences (including emotional text), effectively acting as an annotator. Remarkably, in a cross-lingual setting – projecting English SRL annotations onto Arabic via ChatGPT's translation – then LLM achieved about 94% accuracy, approaching human quality [26].

In conclusion, the period 2020–2025 has seen rapid progress in both Arabic-specific NLP models and the use of large generative models, and the capabilities lie in combining these advances to develop robust, efficient annotation methodologies that leverage the best of human and machine intelligence. The literature review suggests that while LLMs offer valuable annotation support, their optimal use in Arabic NLP requires deeper investigation into prompt engineering, dialect adaptation, and comparison with human oversight. This study builds on these insights by systematically comparing human annotations, CAMeLBERT-DA predictions, and GPT-4o zero-shot outputs on Arabic tweets.

II. Methodology and Experimental Setup

This study explores whether AI models with zero-shot strategies can be used to build the ground truth. It also compares the results using the Cohen Kappa metric and the Pearson correlation coefficient. The analysis of humans vs. AI models in NLP tasks, specifically sentiment analysis, using two Saudi dialect datasets. Fig. 1 illustrates the steps in the general methodology.

A. Datasets

The following is a description of the datasets included in this study's experiments:

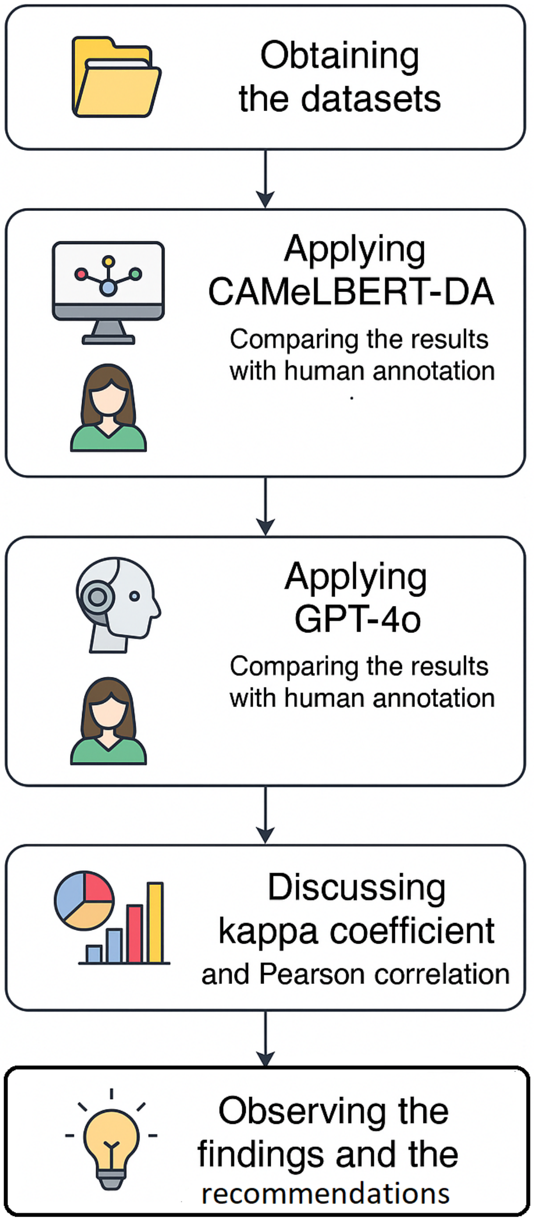

Fig. 1. Methodology overview.

1) SDCT: Multi-Dialects Corpus Classification for Saudi Tweets in [27]. It is dedicated to Saudi dialects. The data are from the X platform, as it is a rich source of data for short posts to be collected. The authors performed the pre-processing phases that involved main tasks, which are data cleaning and normalization. After data collection and pre-processing, the data were manually annotated by human experts with a total of 4181 records. In this study, we used the version that had removed advertisement tweets in the data cleaning phase. Also, it is prepared by removing links (http://), mentions (@Username), retweets, hashtags, and punctuation. While emojis were kept to enrich the sentiment detection.

2) Tb-SAC: Topic-based and Sentiment Classification for Saudi Dialects Tweets. It is extracted from corpora from X, especially from Saudi dialects. The corpus contains 4301 tweets, which are labeled based on sentiments using a common scale:

positive, negative, and neutral. First, preprocessing step by removing all the noise data in the tweets that were not related, such as mentions, hashtags, URLs, white space, retweets, punctuation marks, and redundant tweets. In the second step, normalization is the process of standardizing the form of some Arabic letters that have various shapes to be unified in one, but maintain the same meaning. The dataset was normalized by applying the Tashaphyne library and lemmatization. Emojis here are also removed; only the root words text was kept [28].

B. Models Selection

Based on the NLP task included in the experiment, which is sentiment analysis, the models used are selected. The selected transformer models have two criteria: fine-tuned and have achieved high accuracy in the domain of short tweets. And they consider recent research and benchmarks in the Arabic NLP tasks. The Table I below summarizes the top Arabic transformer models for sentiment analysis, dialect identification, which can be used (either as zero-shot classifiers or pre-trained fine-tuned models). All models are available on Hugging Face (HF). We highlight whether each supports zero-shot use or is a fine-tuned classifier, provide benchmark accuracy/F1 notes, and explain why it's recommended. CAMeL Lab's CAMeLBERT-DA Sentiment Analysis model, fine-tuned for Arabic sentiment classification (positive/negative). This model handles Modern Standard Arabic and dialectal content, as it was trained on diverse Twitter datasets (ASTD, ArSAS, SemEval). (Hugging Face: https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment) CAMeLBERT is a collection of BERT models pre-trained on Arabic texts with different sizes and variants. the release of pre-trained language models for Modern Standard Arabic (MSA), dialectal Arabic (DA), and classical Arabic (CA), in addition to a model pre-trained on a mix of the three [18]. Table I presents the reasons this study adopts this transformer, as many works support [29] and [30].

ChatGPT is an application utilizing the large artificial intelligence models GPT-3 or GPT-4 to generate text based on user input. GPT denotes a neural network adept at learning from extensive datasets to generate natural language outputs. GPT-4 is characterized as more creative, reliable, and collaborative than its predecessor GPT-3 [31] and [32]. The cutting-edge advancement in LLMs [33], GPT-4o was chosen for the comparison in the study experiment via an OpenAI API account for paying customers. It delivers state-of-the-art sentiment classification accuracy without dedicated task training, demonstrating competitive performance against domain-tuned transformer models. Zero-Shot Versatility and Ease of Use: GPT-4 is a large pre-trained model that can perform sentiment analysis without the need for task-specific fine-tuning or large labeled datasets. Broad Language Understanding & Nuance: As a massive foundation model, GPT-4 was trained on an extremely large and diverse corpus (including substantial Arabic content). Table II shows a comparison between the Arabic transformer CAMeLBERT-DA and the large language model GPT-4.

Regarding model structure, GPT-4 and CAMeLBERT-DA have different transformer architectures. GPT-4 (Generative Pre-trained Transformer 4) is built as a decoder-only Transformer, meaning it generates text by predicting one token after another (typical for GPT-style models). CAMeLBERT-DA, on the other hand, is based on BERT, which uses a Transformer encoder – it reads the entire input sequence bidirectionally to produce a contextualized representation. In practice, this means GPT-4 is designed for text generation (and can handle dialogue or any generative tasks), while CAMeLBERT-DA's architecture is designed for understanding text and classification (it outputs encoded features that are fed to a classifier layer) [34] and [35].

C. Evaluation

During the annotation process to build the ground truth for both datasets, human volunteers were involved. In order to validate the human opinions of the sentiment and the AI-generated sentiment, two metrics:

- Inter-model agreement metrics are also used to assess the reliability between annotators. The degree of agreement indicates how accurate the emotion labels assigned to the text. The calculation of the Kappa coefficient and its interpretation were based on the well-known reliability equations and measurements presented in [25]. The kappa statistic revealed the actual reliability of the categorical labels, accounting for the possibility of random agreement.
- Recent NLP studies have employed Pearson correlation to quantify annotator similarity on sentiment and emotion tasks. [36]. Pearson's r is a convenient metric to assess how consistently they apply the scale. (Notably, even annotator confidence scores – if each annotator gives a confidence level for each label.

In summary, Pearson correlation is very useful when treating labels as interval data – it captures how well annotators' scoring trends align – but it should be paired with traditional inter-rater agreement metrics to fully evaluate agreement on the actual labels and to ensure that high correlation indeed reflects reliable, reproducible annotation.

D. Results and Discussion

The experiment here went through the zero-shot settings for both models compared to the human annotators.

1) Generating sentiment using CAMeLBERT-DA transformer: First, using zero-shot and direct inference of the CAMeLBERT transformer to predict the sentiment for dataset1 and dataset2, in Hugging Face [37].

We need to explore the labels produced using the transformer-based model and those in the ground truth by human experts. A detailed comparison from Fig. 2 , the actual sentiment distribution is: negative: 901 (21.5%), neutral: 2506 (59.9%) and positive: 774 (18.5%). on the other hand, Predicted Sentiment Distribution by the transformer was negative: 2607 (62.4%), neutral: 601 (14.4%), and positive: 973 (23.3%). The inter-agreement using Cohen's Kappa Score: 0.2130, while the Pearson Correlation: 0.437. This indicates "fair agreement" between the actual and predicted sentiments. The score suggests that the prediction model performs better than

TABLE I. COMPARISON OF CAMELBERT-DA WITH OTHER ARABIC TRANSFORMERS FOR ARABIC SENTIMENT ANALYSIS

Aspect	**CAMeLBERT-DA**	**Arabic T5 / MARBERT / Others**	**Recommendation**
Accuracy on Arabic Tweets	˜92–93% accuracy on ArSAS, ASTD, and SemEval; fine-tuned on multiple Arabic tweet datasets	MARBERT: competitive but slightly lower; Arabic T5 not specialized or benchmarked for sentiment on tweets	CAMeLBERT-DA outperforms or matches best models due to targeted fine-tuning on diverse tweet datasets
Real-Time Inference Speed	BERT-base encoder model (110M params); efficient for batch and streaming inference	Arabic T5: large encoder–decoder; slower generation; MARBERT: similar size but more general-purpose	CAMeLBERT-DA offers faster, lower-latency inference than T5-like models; ideal for live tweet analysis
Suitability for Tweets	Pretrained on MSA + Dialectal Arabic + social media text; robust to slang, emojis, informal spelling	Arabic T5 trained more on formal MSA and QA-style tasks; MARBERT strong on dialects but not sentiment-specific	CAMeLBERT-DA is dialect-aware, slang-aware, and sentiment-optimized—making it perfect for social content

TABLE II. COMPARISON BETWEEN GPT-4 (CHATGPT) AND CAMELBERT-DA FOR ARABIC SENTIMENT ANALYSIS

Aspect	**GPT-4 (ChatGPT)**	**CAMeLBERT-DA (Arabic BERT)**
Model Type	Large Language Model (LLM); generative, decoder-only transformer; predicts next tokens (autoregressive).	Pretrained BERT-base encoder; bidirectional transformer trained for understanding Arabic. Fine-tuned for classification.
Scale (Parameters)	Very large (estimated 1T+ tokens and hundreds of billions of parameters; exact size not disclosed). Trained on multilingual data including Arabic.	Base-sized model (110M parameters). Pretrained on 5.8B Arabic words from dialectal, MSA, and classical Arabic sources.
Pretraining Data	Trained on diverse multilingual web data (books, articles, web pages). Strong in MSA and general Arabic.	Trained on large-scale Arabic corpora from social media and diverse dialects (CAMeL Lab).
Fine-Tuning for Sentiment Analysis	Not required. Performs sentiment classification in zero- or few-shot mode via prompting. No task-specific finetuning.	Required and completed. Fine-tuned on ASTD, ArSAS, and SemEval datasets. Dedicated sentiment prediction head.
Output Type	Text generation. Can provide label and reasoning if prompted. Very flexible output.	Class label output (e.g., "positive" / "LABEL_1") with probability score. Not designed for explanatory responses.
Sentiment Accuracy on Arabic Tweets	F1-score ≈ 0.75–0.80 (zero- or few-shot). Competitive with fine-tuned models. Excels in MSA and structured dialectal text.	F1-score ≈ 0.72–0.76 (fine-tuned). High accuracy on dialectal Arabic tweets due to domain-specific tuning.
Zero/Few-Shot Support	Native. Easily supports both modes without additional data. Highly flexible.	Not applicable. Requires fine-tuning and training on task-specific data.
Best Use Case	Real-time, on-demand analysis in flexible environments (e.g., chat, APIs). Great for label and explanation.	High-throughput classification on Arabic social media data. Ideal for static datasets with known structure.

random chance, but there's significant room for improvement. Correlation: 0.4377 - There's a moderate positive correlation between the two columns, indicating some relationship but not a strong one.

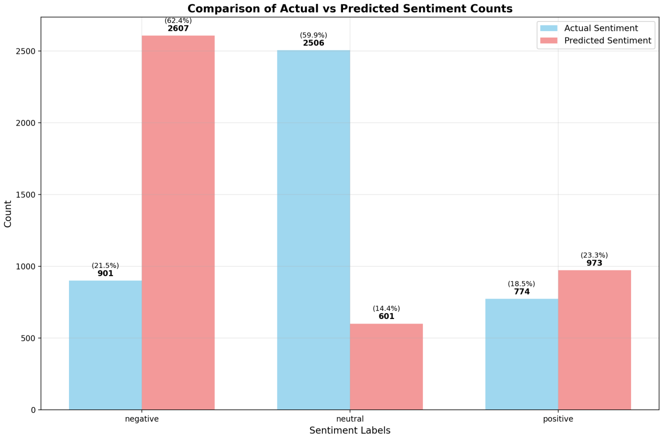

Fig. 2. Comparison of Humans vs. Predicted sentiment counts: (Dataset1).

Key observations:

- The actual sentiment data by humans is predominantly neutral (59.9%), while the predicted model tends to classify more samples as negative (62.4%)
- The model significantly under-predicts neutral sentiment (14.4% vs. 59.9% actual)
- The model over-predicts negative sentiment (62.4% vs. 21.5% actual)
- Positive sentiment predictions are closer to actual values (23.3% vs. 18.5%)

The confusion matrix in Fig. 3 shows where the model struggles most, particularly in correctly identifying neutral sentiments, often misclassifying them as negative. This suggests that positive sentiment is the most clear for humans and AI to detect and identify.

Now, the experiment of implementing the transformer for the second dataset. Key observations are found and perfect Alignment in some cases. The visualization in Fig. 4 shows that when the model is confident about its predictions, there's perfect alignment between feeling and confidence labels. However, the presence of "unknown" labels for negative sentiments indicates areas where the model struggles with certainty, particularly for negative sentiment classification.

Key Insights from the Visualization:

- Perfect Prediction Accuracy for positive and neutral sentiments:
- All 2,086 actual positive sentiments were correctly predicted as positive.

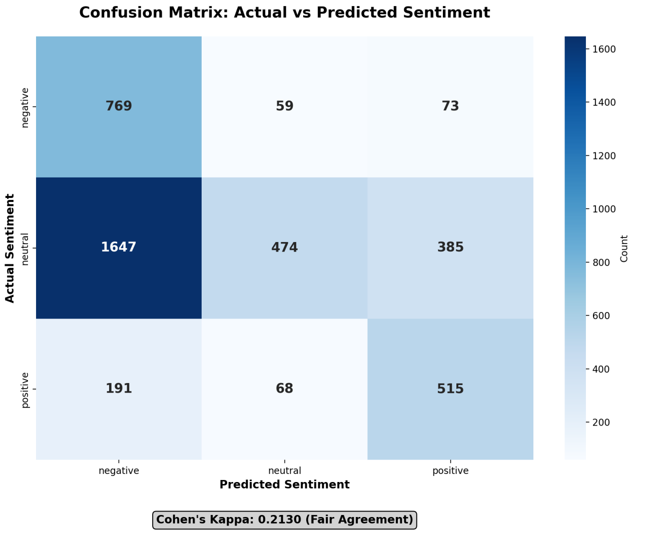

Fig. 3. Confusion matrix of Humans vs. Predicted sentiment labels with cohen's kappa score: (Dataset1).

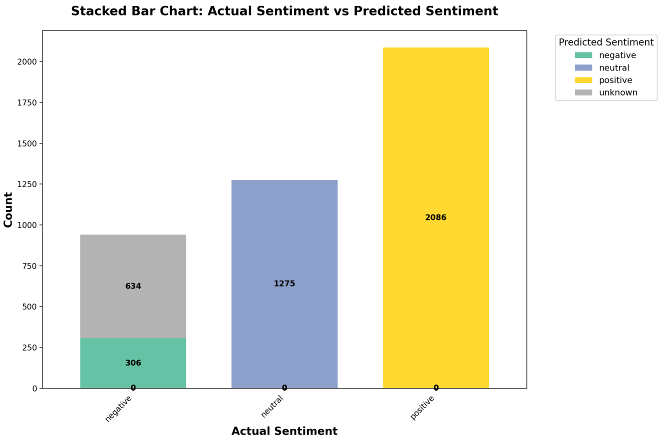

Fig. 4. Stacked bar chart of humans Sentiment vs. Predicted sentiment: (Dataset2).

- All 1,275 actual neutral sentiments were correctly predicted as neutral.
- Mixed Results for Negative Sentiment: Only 306 out of 940 actual negative sentiments were correctly predicted as negative 634 actual negative sentiments were predicted as "unknown", indicating model uncertainty.

The transformer here shows high confidence and accuracy for positive and neutral sentiments but struggles with negative sentiment classification, often defaulting to "unknown" when uncertain. The confusion matrix in Fig. 5 shows that the correlation coefficient between actual humans' sentiment and transformer prediction is 0.114, indicating a weak positive correlation. This suggests that while there is some relationship between the predicted feelings and confidence labels, it's not very strong.

2) GPT-4o: Using the chosen LLM from OpenAI represented in GPT-4o in this part of the experiment to validate the capacity of humans against AI. First of all, a clear, direct prompt based on Zero-Shot Prompt (No examples provided) is prepared as follows:

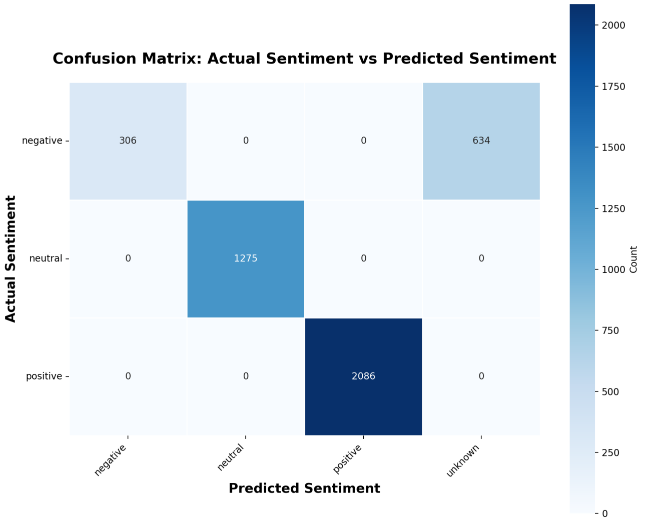

Fig. 5. Confusion matrix of actual Sentiment vs. Predicted sentiment including unknown class: (Dataset2).

Analyze the following text for sentiment, and respond using only one of the three words: Positive, Negative, or Neutral.

Text: 'The service was excellent and very fast.'

Answer:

Positive

We analyzed the agreement of GPT-4o performance against human judgment for both dataset1 and dataset2. Table III shows the sentiment predicted labels using GPT-4o compared to human annotators in dataset1. The visualization shows GPT-4o has fair overall agreement (0.366) with humans but performs inconsistently across sentiment classes - excellent on neutral (81.4%), moderate on positive (63.2%), but poor on negative (39.3%) sentiment.

TABLE III. SENTIMENT ANNOTATION EVALUATION METRICS: GPT-4O VS. HUMANS (DATASET1)

Metric	Value	Interpretation
Cohen's Kappa	0.366	Fair agreement
Overall Agreement	58.0%	Moderate agreement
Pearson Correlation	0.557	Moderate correlation
Negative Agreement	39.3%	Poor agreement
Neutral Agreement	81.4%	Good agreement

On the other hand, an experiment using the second dataset was conducted. Table IV Inter-rater Agreement: The Cohen's Kappa score of 0.239 indicates a fair agreement between the human-labeled and the 'GPT-4o' sentiment classifications. This suggests that while there is some consistency, a notable degree of disagreement also exists. Overall Matching Discrepancy: Only 47.22% of the instances showed an exact match in sentiment labels between the two sources, highlighting a significant divergence in their overall classifications. Correlation

TABLE IV. SENTIMENT ANNOTATION EVALUATION METRICS: GPT-4O VS HUMANS (DATASET2)

Metric	Value	Interpretation
Cohen's Kappa	0.239	Fair agreement
Overall Agreement	47.22	Poor agreement
Pearson Correlation	0.433	Moderate correlation
Negative Agreement	54.46	Moderate agreement
Neutral Agreement	43.29	Poor agreement
Positive Agreement	45.60	Poor agreement

Strength: A moderate positive correlation (0.45) was observed between the numerical representations of the actual humans' feelings and 'GPT-4o' sentiments. This implies a general tendency for them to align, but not a strong linear relationship. Sentiment-Specific Consistency: Agreement levels vary across sentiment categories. Negative sentiment classifications show the highest specific agreement at 54.46%, suggesting better consistency for negative cases. Conversely, neutral sentiment exhibits the lowest specific agreement at 43.29%, indicating more challenges in aligning neutral classifications.

Finally, observations during the comparisons are as follows:

- The results show closer opinions of LLMs when the zero-shot setting is implemented and are more correlated to human experts' judgments. Transformers tend to show a moderate correlation with human annotators, while GPT-4o shows fair agreement. This is due to the fine-tuning setting of transformers for the specified domain to adapt the dataset.
- The efficiency and the cost of time using OpenAI for developers took 120 minutes to generate sentiment for every dataset, which counts around 4000 posts based on Google Colab using a paid account. On the other hand, the transformer CAMeLBERT-DA offline pipeline took only 60 seconds to generate the sentiment. This is because of the broad general use of LLMs and complexity of models and the number of parameters compared to the transformer, which is usually domain-specific.
- Another observation is the dataset level of preprocessing affects the AI decisions. In other words, during the experiment, the dataset2 version that had an extra level of preprocessing via lemmatization and removing all the emojis reduced the capability of models to detect the right sentiment, as it provides the lowest inter-agreement between GPT-4o (0.24) and CAMeLBERT-DA (0.14).

III. CONCLUSION AND FUTURE WORK

In this research, we compared OpenAI's GPT-4o model and transformer-based model against human experts for the annotation process. The experiments were conducted for NLP tasks in specific sentiment analysis prediction for Saudi dialect tweets. The finding revealed a better performance of GPT-4o against the transformer CAMeLBERT-DA. Inter-agreement Cohen Kappa was 0.366, indicating fair agreement between GPT-4o and human annotation for the first dataset, where initial preprocessing is conducted without any lemmatization, and emojis also remain in the text. Notably, datasets with intensive preprocessing produced difficulties in sentiment detection for transformers. This suggests that transformers and LLMs require more context to generate sentiment close to humans. The comparison between the models' performance and the human annotation explains that zero-shot manner in this experiment is fair, but not a completely reliable tool to substitute the human annotation process. For these reasons, sentiment generation research should investigate the combination of human crowdsourcing and automatic sentiment generation by LLMs. Furthermore, a limitation of this study is the lack of comparison across different Arabic dialects, such as Egyptian, Gulf, and Maghrebi. Also, in the future, employing few-shot strategies for LLMs and fine-tuned transformers can be studied for sentiment tasks and other NLP tasks such as Arabic text summarization. New evaluation metrics and empirical studies are critical to increasing reliability and trustworthiness when zero-shot methods are applied in real-time applications. In addition, the implications of fairness and bias issues can also be studied, which are crucial for ethical real-world deployment.

REFERENCES

[1] A. Elmadany, E. M. B. Nagoudi, and M. Abdul-Mageed, "Orca: A challenging benchmark for arabic language understanding," *arXiv preprint arXiv:2212.10758*, 2022.

[2] S. Aggarwal, B. Khanna, R. Madhumala, R. Garg, A. Maroju, and B. Brahma, "Analysis of arabic texts using semantics," in *International Conference on Computing and Communication Networks*. Springer, 2024, pp. 389–401.

[3] J. Algaraady and M. Mahyoob, "Exploring chatgpt's potential for augmenting post-editing in machine translation across multiple domains: challenges and opportunities," *Frontiers in Artificial Intelligence*, vol. 8, p. 1526293, 2025.

[4] G. Bourahouat, M. Abourezq, and N. Daoudi, "Systematic review of the arabic natural language processing: challenges, techniques and new trends," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 3, pp. 1333–1343, 2023.

[5] Z. Alyafeai, M. S. Alshaibani, B. AlKhamissi, H. Luqman, E. Alareqi, and A. Fadel, "Taqyim: Evaluating arabic nlp tasks using chatgpt models," *arXiv preprint arXiv:2306.16322*, 2023.

[6] M. Abdul-Mageed, A. Elmadany, and E. Nagoudi, "Arbert & marbert: Deep bidirectional transformers for arabic," *arXiv preprint arXiv:2101.01785*, 2020.

[7] A. Bourahouat *et al.*, "A survey on resources and tools for arabic nlp," *ACM Computing Surveys*, 2023.

[8] N. Habash *et al.*, "The state of arabic nlp: Resources and challenges," *Journal of Computational Linguistics*, 2022.

[9] B. M. Almotairy, M. Abdullah, and D. H. Alahmadi, "Dataset for detecting and characterizing arab computation propaganda on x," *Data in Brief*, vol. 53, p. 110089, 2024.

[10] A. Alharbi *et al.*, "Challenges in annotating dialectal arabic: An empirical study," in *Proceedings of LREC*, 2021.

[11] W. Zaghouani and M. R. Biswas, "An annotated corpus of arabic tweets for hate speech analysis," *arXiv preprint arXiv:2505.11969*, 2025.

[12] R. Daud, N. Basir, N. F. N. Mohd Rafei Heng, M. M. S. Meor Sepli, and M. Melinda, "Large language model versus human: A study of automatic data annotation in islamophobia dataset," *Available at SSRN 5045167*, 2024.

[13] A. S. Alzahrani, D. H. Alahmadi, N. M. Alharbi, and H. A. Almagrabi, "Blockchain-based crowdsourcing framework for machine learning ground truth," *IEEE Access*, 2025.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[15] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, 2020.

[16] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)*, 2020, pp. 9–15.

[17] A. Safaya, M. Abdullatif, and J. Ren, "Arabic-bert: Benchmarking pretrained language models for arabic," *arXiv preprint arXiv:2003.00104*, 2020.

[18] G. Inoue, N. Habash, D. Taji, H. Bouamor, and W. Zaghouani, "Camelbert: Transformer-based arabic language models pretrained on diverse corpora," *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pp. 91–103, 2021. [Online]. Available: https://aclanthology.org/2021.wanlp-1.10

[19] L. Xue *et al.*, "mt5: A massively multilingual pretrained text-to-text transformer," in *NAACL*, 2021.

[20] X. V. Lin *et al.*, "Few-shot learning with multilingual language models," *Transactions of the ACL*, 2022.

[21] OpenAI, "Gpt-4 technical report," 2023, https://openai.com/research/gpt-4.

[22] S. Khalifa *et al.*, "Evaluating chatgpt and gpt-4 for arabic tasks: A case study in sentiment and sarcasm," *arXiv preprint arXiv:2305.12345*, 2023.

[23] A. A. Alsuwaylimi and Z. S. Alenezi, "Leveraging transformers for detection of arabic cyberbullying on social media: Hybrid arabic transformers." *Computers, Materials & Continua*, vol. 83, no. 2, 2025.

[24] M. I. Khondaker, M. Mahmoud, M. Torres, and A. Nazarenko, "Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp tasks," *arXiv preprint arXiv:2308.08794*, 2023.

[25] S. Wang *et al.*, "Human-ai collaboration in data labeling: Benefits and pitfalls," in *Proceedings of ACL*, 2022.

[26] F. Senator, A. Lakhfif, I. Zenbout, H. Boutouta, and C. Mediani, "Leveraging chatgpt for enhancing arabic nlp: Application for semantic role labeling and cross-lingual annotation projection," *IEEE Access*, 2025.

[27] A. Bayazed, O. Torabah, R. AlSulami, D. Alahmadi, A. Babour, and K. Saeedi, "Sdct: multi-dialects corpus classification for saudi tweets," *methodology*, vol. 11, no. 11, 2020.

[28] S. Alzahrani, F. Alruwaili, D. Alahmadi, and K. Saeedi, "Tb-SAC: Topic-based and Sentiment Classification for Saudi Dialects Tweets," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 20, no. 9, pp. 41–49, 2020. [Online]. Available: http://ijcsns.org/07_book/html/202009/202009006.html

[29] A. Alsudais, "Comparing deep learning arabic sentiment analysis models using arsas dataset," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 74–81, 2023.

[30] S. Alhumoud and R. Alhajri, "Arabic sentiment analysis on twitter: A comparative study using deep learning and pretrained transformers," *Applied Sciences*, vol. 13, no. 5, p. 2741, 2023.

[31] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu *et al.*, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-radiology*, vol. 1, no. 2, p. 100017, 2023.

[32] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "Chatgpt: Applications, opportunities, and threats," in *2023 systems and information engineering design symposium (SIEDS)*. IEEE, 2023, pp. 274–279.

[33] R. Islam and O. M. Moushi, "Gpt-4o: The cutting-edge advancement in multimodal llm," *Authorea Preprints*, 2024.

[34] M. Hannani, A. Soudi, and K. Van Laerhoven, "Assessing the performance of chatgpt-4, fine-tuned bert and traditional ml models on moroccan arabic sentiment analysis," in *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 2024, pp. 489–498.

[35] M. Saranya and B. Amutha, "A comparative study on the evaluation of chatgpt and bert in the development of text classification systems," in *Generative Artificial Intelligence and Ethics: Standards, Guidelines, and Best Practices*. IGI Global, 2025, pp. 91–108.

[36] J. A. Lossio-Ventura, R. Weger, A. Y. Lee, E. P. Guinee, J. Chung, L. Atlas, E. Linos, and F. Pereira, "A comparison of chatgpt and fine-tuned open pre-trained transformers (opt) against widely used sentiment analysis tools: sentiment analysis of covid-19 survey data," *JMIR Mental Health*, vol. 11, p. e50150, 2024.

[37] CAMeL Lab, "bert-base-arabic-camelbert-da-sentiment," https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment, accessed: 2025-07-21.