

Collaborative Emotion Annotation: Assessing the Intersection of Human and AI Performance with GPT Models

Hande Aka Uymaz^a and Senem Kumova Metin^b
İzmir University of Economics, Department of Software Engineering, İzmir, Turkey

Keywords: Emotion, Sentiment, Lexicon, Annotation, Cohen's Kappa, Fleiss Kappa.


Abstract: In this study, we explore emotion detection in text, a complex yet vital aspect of human communication. Our focus is on the formation of an annotated dataset, a task that often presents difficulties due to factors such as reliability, time, and consistency. We propose an alternative approach by employing artificial intelligence (AI) models as potential annotators, or as augmentations to human annotators. Specifically, we utilize ChatGPT, an AI language model developed by OpenAI. We use its latest versions, GPT3.5 and GPT4, to label a Turkish dataset having 8290 terms according to Plutchik's emotion categories, alongside three human annotators. We conduct experiments to assess the AI's annotation capabilities both independently and in conjunction with human annotators. We measure inter-rater agreement using Cohen's Kappa, Fleiss Kappa, and percent agreement metrics across varying emotion categorizations- eight, four, and binary. Particularly, when we filtered out the terms where the AI models were indecisive, it was found that including AI models in the annotation process was successful in increasing inter-annotator agreement. Our findings suggest that, the integration of AI models in the emotion annotation process holds the potential to enhance efficiency, reduce the time of lexicon development and thereby advance the field of emotion/sentiment analysis.


1 INTRODUCTION

Emotion and sentiment are two terms that define the feelings of people. Emotion encompasses a broad range of distinct categories, such as happiness, anger, fear, and sadness, among others. In contrast, sentiment is a more general feeling that can be categorized as positive, negative, or neutral. People may experience different emotions or sentiments in the same circumstances. This can be affected by a variety of factors, including gender, age, psychology, culture, and personal experiences. Considering the differences between emotions and sentiments, emotion detection becomes a challenging task. Emotion and sentiment play a crucial role in human communication and expression. Text, speech, video, or EEG signals are used to reflect the emotions of people. For centuries, texts have served as a means of communication and have consequently become a valuable source for studies on emotion detection. Furthermore, with the popular usage of social media, collecting data for emotion analysis studies is easier.

There are several ways to extract emotive data from a data source. For instance, we naturally remark facial expressions, body language, tone of voice, and gestures as powerful indicators of emotions. Word choice, frequency of word usage, sentence structure, use of emoticons and emojis, and context are among the key factors examined in text-based emotion detection.

There exist multiple approaches to determining the emotion or sentiment conveyed in a text, including machine learning techniques and lexicon-based methods. Having an emotion/sentiment lexicon, which is a list of labelled words, is generally a necessity to apply these methods. There are multiple methods for collecting annotated datasets, including expert, crowd and automated annotation (e.g., Staiano & Guerini, 2014; Schuff et al., 2017; Mohammad & Turney, 2010; Aka Uymaz & Kumova Metin, 2022)). Each method has its own advantages and disadvantages in terms of reliability, time, and consistency. For example, expert annotation involves the expertise of domain experts or linguists, but it can

^a  <https://orcid.org/0000-0002-3535-3696>

^b  <https://orcid.org/0000-0002-9606-3625>

be a time-consuming and costly process. Moreover, examining the previous studies, the lexicons proposed are mainly in English, which has the most resources. The research on languages other than English utilizes translated versions or constructs the lexicon from the ground up. Utilizing translated versions for creating emotion/sentiment lexicons poses several challenges. A significant concern is the loss of nuances and cultural distinctions, as emotions are intricately tied to cultural contexts, and direct translations may not fully capture their intricacies. Moreover, some languages lack equivalent words, leading to approximate matches and potential inaccuracies. Ambiguity and polysemy in words across languages further complicate matters, as a single word can carry multiple meanings and interpretations. Additionally, languages evolve over time, introducing new emotion-related expressions that may not be adequately reflected in translated lexicons. Keeping such lexicons up-to-date requires frequent revisions. Due to such reasons, it would be advantageous to develop emotion/sentiment lexicons natively in the target language, through collaboration with linguists and native speakers, ensuring the accurate representation of cultural and contextual nuances of emotions.

In this study, considering the difficulties of forming an annotated dataset, we considered utilizing AI models as an alternative to human annotators or increasing the number of annotators by forming a combination with human annotators.

As an artificial intelligence, we utilize the ChatGPT AI language model (OpenAI, 2023) developed by OpenAI. ChatGPT is trained on massive amounts of data to understand and generate text-based outputs, like humans. We employ its latest versions, GPT3.5 and GPT4, as two different AI annotators for our emotion labeling task. We also have three human annotators, and they, along with two AI language models, labeled the same Turkish dataset according to Plutchik's eight emotion categories and the neutral category. We carried out multiple experiments to assess the annotation capabilities of AI models, both with and without the involvement of human annotators.

We measured the inter-rater agreement among annotators in different perspectives employing three metrics which are Cohen's Kappa, Fleiss Kappa and percent agreement. These statistics are evaluated when lexicon words are labeled according to eight, four or binary emotion categories. We utilized Plutchick's emotion categories for eight discrete emotion categories (Robert Plutchik, 1980). Then, we eliminated them to 4 basic emotions. For binary

emotion labeling we considered the label of a word as an emotive word or non-emotive word.

The remainder of the paper is structured as follows. Section II provides an overview of previous work related to the topic. Section III describes the process of lexicon annotation employed in this study. In Section IV, we present the experimental results obtained from our analysis. Finally, Section V gives the conclusion.

2 LITERATURE REVIEW

Emotion or sentiment lexicons are linguistic resources, where each entry consists of a word associated with its respective emotional or sentiment label/labels. Typically, sentiment lexicons are structured around polarity values, assigning words to categories of positive, negative, or neutral. On the other hand, emotion lexicons consist of words with discrete emotion categories (e.g., happiness, sadness, or anger) or emotional dimensions such as pleasure and arousal.

The process of annotating lexicons and the quality of annotation are vital as they directly impact the performance of emotion detection studies. Annotators need to label independently from each other by following a specific guideline. Different methods exist for the collection of annotated datasets: expert annotation, crowd-sourced annotation, and automated annotation. For example, National Research Council Canada (NRC) emotion lexicon (Mohammad & Turney, 2013) is a frequently utilized publicly available emotion lexicon. It is based on English, and there exists automatically translated versions for other languages. The lexicon is constructed by crowdsourcing annotation with Amazon's Mechanical Turk service. While crowdsourcing offers several benefits such as cost-effectiveness and speedy turnaround times, it also presents certain drawbacks. For example, the quality of annotation can be a concern, with issues like random or incorrect annotations and not following all the directions for the annotation process. Thus, the educational background of participants remains uncertain in the crowdsourcing environment (Mohammad & Turney, 2013). DepecheMood is another emotion lexicon where automatic annotation was carried out by gathering data with the extraction of news articles considering readers' selection of emotion categories related to the news (Staiano & Guerini, 2014). Valence Aware Dictionary for sEntiment Reasoning (VADER) (Hutto & Gilbert, 2014) is a sentiment polarity lexicon annotated by expert human

annotators. To ensure the collection of meaningful data, certain quality control measures were implemented for the annotators, which included testing for English comprehension and sentiment rating abilities.

Although there are more studies on English dataset construction, as in other natural language processing tasks, a limited number of emotion lexicons/datasets have been developed in different languages. TEL lexicon (Toçoglu & Alpkoçak, 2019), which is valuable as an alternative to the lexicon that is the output of our study, is one of them. TEL is the first constructed Turkish emotion lexicon based on TREMO dataset (Tocoglu & Alpkocak, 2018) which has 27,350 documents collected from 4709 people. Furthermore, the dataset has been validated by 48 annotators to address ambiguities in documents. It has several versions based on different enrichment methods and preprocessing techniques that have 3976 to 7289 terms. Gala and Brun (2012) proposed a French sentiment lexicon having 7483 nouns, verbs, adjectives and adverbs. In the lexicon, terms are annotated by polarity values are labeled by a semi-automatic method. Navarrete et.al., presented a Spanish emotion lexicon (Segura Navarrete et al., 2021). The Emotion Lexicon includes specific emotions and their corresponding intensities associated with 1892 Spanish words. EmoLex lexicon (Mohammad & Turney, 2013) has been translated, validated, and further enhanced with synonyms derived from WordNet (Strapparava & Valitutti, 2004).

In the work of Aka Uymaz and Kumova Metin (2022) a detailed review on emotion lexicons and other resources employed in emotion/sentiment detection is provided.

AI models such as ChatGPT and other advanced language models have brought about a paradigm shift in Natural Language Processing (NLP) due to their exceptional capabilities. These models, driven by advanced deep learning algorithms and trained on extensive textual data, perform exceptionally well in comprehending and generating human-like text. The possibilities for their use in NLP are vast and promising. One key application involves virtual assistants, which can now engage in more natural and contextually relevant conversations, leading to improved customer support and assistance (Day & Shaw, 2021). Furthermore, these models have revolutionized sentiment analysis, empowering businesses to accurately assess customer feedback. For instance, Kertkeidkachorn and Shira introduced an approach that combines graph neural networks with a model called Graph Neural Network-based

model with the pre-trained Language Model (Kertkeidkachorn & Shirai, n.d.). The model effectively captures the connection between users and products by generating distributed representations through a graph neural network. These representations are then integrated with distributed representations of reviews from the RoBERTa (Liu et al., 2019) pre-trained language model to predict sentiment labels. Additionally, language models have also simplified language translation, making cross-lingual communication seamless. Furthermore, they play a pivotal role in content creation, summarization, and recommendation systems, enabling users to access relevant information more efficiently. For example, the survey of Cao et.al. explores the growing interest in Generative AI techniques, such as ChatGPT (OpenAI, 2023), DALL-E-2 (Ramesh et al., 2021), and Codex (Chen et al., 2021), and their application in Artificial Intelligence Generated Content that involves the efficient production of digital content using AI models based on human instructions and intent information (Cao et al., 2023). The integration of large-scale models has enhanced the extraction of intent and content generation, leading to significantly more realistic results.

In this study, we utilized the ChatGPT language model with both GPT3.5 and GPT4 architectures to annotate the emotion lexicon in the Turkish language. Although this language model has been used in various natural language processing tasks, as far as we know, it has not been studied for lexicon annotation and the Turkish language yet.

3 LEXICON ANNOTATION

In this study, we construct a Turkish emotion lexicon based on 8 discrete categories of Plutchik's theory of emotions which are joy, sadness, anger, fear, trust, disgust surprise and anticipation (Robert Plutchik, 1980). The words for our lexicon have been sourced from the frequently used English NRC emotion lexicon (Mohammad & Turney, 2013). Totally, there are 8290 terms are annotated for this lexicon. The terms are translated to Turkish manually. Afterwards, the word list was reviewed by the researchers and the observed translation errors were corrected.

For the annotation of emotion labels for these Turkish words, a combination of three human annotators and two AI annotators are employed. The human annotators consist of three individuals who are native Turkish speakers and currently enrolled in a master's degree program who will be named as A1, A2 and A3. We used two models of ChatGPT as artificial

intelligence annotators (GPT3.5 and GPT4). ChatGPT is a language model developed by OpenAI, based on the transformer based GPT (Generative Pretrained Transformer) architecture (Radford et al., 2019). Both versions, GPT3.5 and GPT4 are trained on vast amounts of internet data and designed to produce answers like human. They can be utilized in a variety of tasks such as, translation, question answering, tagging, or classifying information. It has been trained on a multilingual corpus and can generate text in a variety of languages. GPT3.5 has been launched as the fastest model suitable for a wide range of everyday tasks, while GPT4 is described as a more proficient version designed specifically for tasks that demand creativity and advanced reasoning abilities.

We asked both human annotators and AI annotators to match the given words with a maximum of two out of Plutchick's eight emotion categories or neutral category and specify the primary and secondary emotions, if applicable. During the word labeling of these two AI annotators, we realized that GPT3.5 model sometimes labeled the words with other emotion categories that are not specified in the query. However, in contrast, GPT4 was much more successful in complying with the given instructions.

4 EXPERIMENTAL RESULTS

After completing the entire annotation process, we conducted three analyses to measure inter-rater reliability using three different metrics:

- (1) Cohen's Kappa (Jacob Cohen, 1960),
- (2) Fleiss Kappa (Joseph L. Fleiss, 1971),
- (3) Percent agreement.

These are the metrics that are accepted as reliable measures of inter-rater agreement in various fields (e.g., Atapattu et al., 2022; Pérez et al., 2020; Pham et al., 2023; Wilbur et al., 2006).

Cohen's Kappa is a statistical measure used to assess level of agreement between two raters. It quantifies how well two observers' assessments align based on the same conditions. If there is no agreement, its value is less than 0, while perfect agreement is represented by a value of 1. On the other hand, Fleiss Kappa is a metric used to assess the level of agreement among multiple raters or observers. It extends the concept of Cohen's Kappa to situations where there are more than two raters. Fleiss kappa and Cohen's kappa both consider the possibility of agreement occurring by chance. Essentially, these statistical measures evaluate the level of agreement while considering the agreement that could be expected to happen randomly.

Percent agreement is a statistical metric used to assess the level of agreement among multiple annotators based on their categorical decisions. It is calculated by simply dividing the total number of agreements by the total number of samples and representing it as a percentage.

Table 1 presents the number of labels (in percentages) assigned by human and AI annotators. For example, the first column presents that annotator A1 labels 5.1% of all samples with Anger category. Last four columns, in Table 1 show the percentages for majority of votes. The term majority of votes refers to the approach where a sample is labeled with the emotion category which gets the most votes (at least half of the votes) in multi-annotator environment. To exemplify, when all human and AI annotators, totally 5 annotators, are involved (A1-A2-A3-GPT3.5-GPT4) in labeling (last column in Table 1), 3.4% of samples are tagged as Anger by at least 3 annotators. In our study, the samples that are not labeled with the same category by most of the annotators are set as "No label".

In Table 1, the bold cells refer to the label that holds the highest percentage for each annotator, except no label case. Examining the results in Table 1, it is seen that the percentage of words that are labeled as Joy are dominantly higher except A2. It has also been observed that AI models tend to assign approximately equal number of samples to all labels when compared to human annotators.

Table 2 represents the results of Cohen's Kappa and percent agreement measures. The table presents all possible pairwise combinations of annotators. We calculated Cohen's Kappa and percent agreement scores for three versions presented as 8-emotion, 4-emotion, binary (emotive or no-emotive). 8-emotion version refers to the experiment where the labels given by annotators are employed without any preprocessing. In 4-emotion version, samples labeled with emotions other than 4 core emotions (anger, fear, sadness and joy) are assigned to neutral category. And binary version is the case where samples who are labeled with one emotion is accepted to be in emotive group and others are assigned to non-emotive. In Table 2, the three highest values are highlighted in bold in every column.

The experimental results in Table 2 revealed the following outcomes:

- (1) Both 8 and 4-emotion versions, the higher Cohen's Kappa values are obtained with annotator pairs GPT 3.5-GPT 4, A1-A2, and GPT 4 and the majority of votes of 3 human annotators.

- (2) The highest Cohen’s Kappa value is obtained with annotators A1 and A2 when considering only human judges.
- (3) Examining the percent agreement scores, the score between human raters is higher than the other pairs in binary emotion labeling.
- (4) The highest percent agreement is obtained with GPT 3.5–GPT 4 pair considering eight and four emotion categories.

Table 1: Emotion label percentages of annotators (Totally 8290 Samples).

	A1	A2	A3	GPT3.5	GPT4	A1-A2-A3	A1-A2-A3-GPT3.5	A1-A2-A3-GPT4	A1-A2-A3-GPT3.5-GPT4
Anger	5.1%	6.5%	21.3%	3.0%	5,6%	4.7%	5.9%	7.0%	3.4%
Anticipation	16.5%	19.0%	15.3%	4.1%	3,1%	10.7%	12.1%	11.8%	3.2%
Disgust	7.3%	5.4%	2.7%	5.2%	3,6%	2.4%	3.4%	3.2%	1.6%
Fear	10.3%	14.8%	9.2%	4.0%	5,1%	7.2%	8.1%	8.2%	4.1%
Joy	22.3%	22.4%	10.1%	7.9%	9.9%	15.3%	17.0%	17.6%	9.2%
Sadness	12.6%	9.8%	7.6%	3.1%	5,7%	6.5%	6.9%	7.1%	4.5%
Surprise	5.5%	3.6%	11.5%	4.1%	1,4%	2.1%	2.9%	2.3%	0.8%
Trust	10.8%	11.6%	17.6%	6.9%	4,8%	7.4%	8.5%	8.3%	3.7%
Neutral	9.6%	6.8%	4.7%	61.8%	60,8%	0.0%	0.0%	0.0%	0.0%
No label	-	-	-	-	-	39.7%	23.9%	22.5%	57.0%

Table 2: Cohen's Kappa and Percent Agreement scores (The top three values are indicated in bold).

Pair	Cohen’s Kappa			Percent Agreement			
	8-Emotions	4-Emotion	Binary	8-Emotions	4-Emotion	Binary	
Human & Human annotator	A1 - A2	0.303	0.366	0.289	39.88%	56.36%	89.28%
	A1 - A3	0.089	0.101	0.137	18.90%	37.93%	88.49%
	A2 - A3	0.104	0.123	0.246	20.75%	38.43%	91.81%
AI & Human annotator	GPT3.5 - A1	0.150	0.209	0.068	24.01%	55.30%	44.49%
	GPT3.5 - A2	0.129	0.203	0.048	20.84%	52.94%	42.67%
	GPT3.5 - A3	0.059	0.089	0.024	12.83%	49.48%	40.75%
	GPT3.5 - Majority	0.196	0.282	0.066	27.14%	55.78%	46.26%
	GPT4 - A1	0.214	0.326	0.106	30.01%	59.82%	47.45%
	GPT4 - A2	0.197	0.342	0.065	27.27%	59.29%	44.50%
	GPT4 - A3	0.073	0.135	0.025	14.22%	49.20%	41.69%
	GPT4 - Majority	0.290	0.453	0.094	36.22%	64.50%	50.58%
AI & AI annotator	GPT3.5 - GPT4	0.395	0.433	0.447	63.49%	78.35%	73.78%

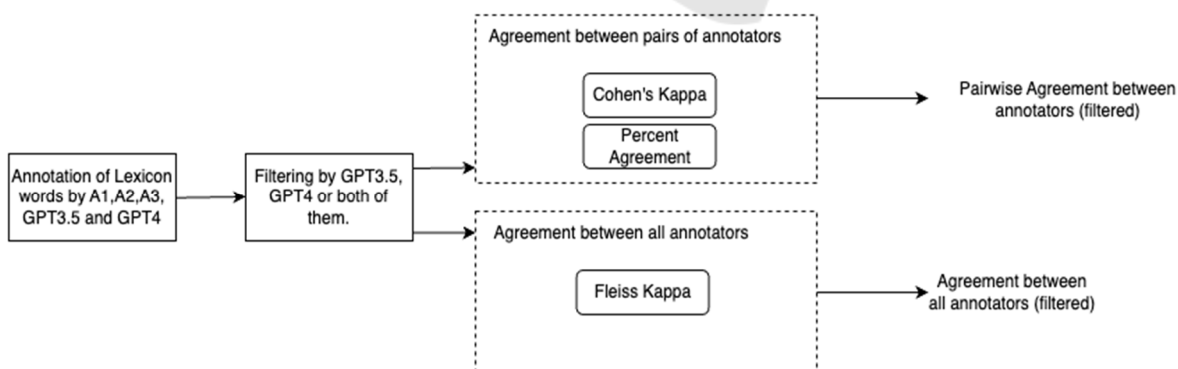


Figure 1: Framework for annotation procedure with filtering.

Table 3: Cohen's Kappa and Percent Agreement scores filtered by GPT 3.5 and GPT 4 (The top three values are indicated in bold).

	Pair	Cohen's Kappa		Percent Agreement	
		8-Emotions	4-Emotion	8-Emotions	4-Emotion
AI (GPT 3) & Human annotator	GPT3.5_Filtered - A1	0.354	0.619	43.96%	72.41%
	GPT3.5_Filtered - A2	0.320	0.617	41.04%	72.28%
	GPT3.5_Filtered - A3	0.146	0.305	24.82%	47.22%
	GPT3.5_Filtered - Majority	0.451	0.724	52.74%	80.11%
AI (GPT 4) & Human annotator	GPT4_Filtered - A1	0.469	0.652	54.73%	74.31%
	GPT4_Filtered - A2	0.469	0.662	55.16%	75.18%
	GPT4_Filtered - A3	0.173	0.300	27.90%	47.64%
	GPT4_Filtered - Majority	0.611	0.754	67.61%	82.04%
AI & AI annotator	GPT3.5_Filtered - GPT4_Filtered	0.533	0.781	59.75%	83.99%

Our experimental results showed us that not only the agreement between AI models and human annotators is low, but also the agreement between humans is low in emotion labeling. Based on this result, we experimented to use the decisions of artificial intelligence models as a filter. The general framework used for filtering procedure is represented in Figure 1. Namely, we eliminate the words labeled as “neutral” by AI annotators and recalculated the Cohen’s Kappa and percent agreement scores as can be seen in Table 3 (highest three values are indicated as bold in every column). By examining the Cohen’s Kappa values, it is clear that the removal of the “neutral” label (filtering) increased the level of agreement between AI models and human annotators for both the 8-emotion and 4-emotion categories compared to the unfiltered data. According to results, both in eight and four emotion categories, the Cohen’s kappa value between both AI models and the majority of votes of three human annotators are higher than the pairs of AI models and a single human annotator.

Tables 4-6 presents Fleiss Kappa statistic results between 3 human annotators or between 3 human annotators and GPT 3.5 or/and GPT 4. As can be seen from the tables, filtering of “neutral” categories again increased the Fleiss Kappa values between all annotator groups (except some cases in binary labelling). Furthermore, examining Table 4, adding GPT3.5 as an annotator increased the Fleiss Kappa value from 0.19 to 0.23, 0.22 to 0.25 and 0.14 to 0.95 with eight, four and binary emotion categories, respectively. The similar improved results are obtained when adding GPT4 as an annotator with human raters (from 0.21 to 0.28, from 0.23 to 0.30 and from 0.10 to 0.97) as can be seen in Table 5.

Finally, Table 6 presents the Fleiss Kappa statistic results of all human and AI annotators. Filtering (eliminating) the terms labeled as “neutral” by GPT3.5 and GPT4 results in an increase in inter annotator agreement when adding AI annotators to the labeling process compared to only expert annotators.

Table 4: Fleiss Kappa scores between human annotators and GPT 3.5.

	8-Emotions	4-Emotions	Binary
A1 - A2 - A3	0.16	0.19	0.22
A1 - A2 - A3 - GPT3.5	0.11	0.17	0.66
A1 - A2 - A3 (Filtered)	0.19	0.22	0.14
A1 - A2 - A3 - GPT3.5 (Filtered)	0.23	0.25	0.95

Table 5: Fleiss Kappa scores between human annotators and GPT 4.

	8-Emotions	4-Emotions	Binary
A1 - A2 - A3	0.16	0.19	0.22
A1 - A2 - A3 - GPT4	0.14	0.22	0.67
A1 - A2 - A3 (Filtered)	0.21	0.23	0.10
A1 - A2 - A3 - GPT4(Filtered)	0.28	0.30	0.97

In Table 7, the distribution of labels when samples that are either labeled as neutral by GPT3.5 or GPT4 are removed from the data set is given. The table provides for two sets. First column (A1-A2-A3) refers to the set that is the filtered version of majority of votes for human annotators. And the second is filtered set of majority of votes for all annotators. In Table 7, it is examined that still the top-most

Table 6: Fleiss Kappa scores between all human and AI annotators.

	8-Emotions	4-Emotions	Binary
A1 - A2 - A3	0.16	0.19	0.22
A1 - A2 - A3 - GPT3.5 - GPT4	0.14	0.21	0.60
A1 - A2 - A3 (Filtered)	0.22	0.23	0.08
A1 - A2 - A3 - GPT3.5 - GPT4(Filtered)	0.31	0.32	0.98

percentage belongs to the Joy category. In addition, four core emotions (joy, anger, fear, and sadness) have dominantly more samples in final lexicon. It is examined that as the filter (GPT3.5+GPT4) is applied from Table 7, the data set size is decreased from 8290 to 2119 samples. But on the other hand, it is observed that the annotation process becomes much more feasible by decreasing the time and effort in human annotation. As a result, it can be stated that an increased number of initial samples may be provided to AI models to determine the emotive samples and only samples determined to contain emotion can be submitted for human annotators to label.

Table 7: Emotion label percentages after filtering (Totally 2119 Samples).

	A1-A2-A3 (Filtered)	A1-A2-A3-GPT3.5-GPT4 (Filtered)
Anger	10,3%	10,3%
Anticipation	4,4%	5,1%
Disgust	4,2%	2,8%
Fear	11,9%	13,8%
Joy	19,8%	17,5%
Sadness	11,3%	11,9%
Surprise	2,4%	2,0%
Trust	6,9%	6,0%
Neutral	0,0%	0,0%
No label	57,0%	30,3%

5 CONCLUSION

This study focuses on creating an annotated dataset that can be used in emotion detection studies. Labeling set of words is a process that often presents challenges due to reliability, time, and consistency.

We explored an alternative approach using AI models, specifically ChatGPT versions GPT3.5 and GPT4, as annotators for a Turkish dataset with 8290 terms, based on Plutchik’s eight emotion categories. Using three human annotators, we conducted experiments to assess the AI’s annotation capabilities independently and in combination with human annotators. The experiments are performed over not

only 8 emotion labeled version of the dataset but also on the versions where four core emotions and emotive/non-emotive labels are considered.

By measuring inter-rater agreement using Cohen’s Kappa, Fleiss Kappa, and percent agreement metrics, we found that integrating AI models in the annotation process increased inter-annotator agreement. Especially, when the AI decision is employed as a preprocessing filter, the agreement among annotators comes to an acceptable level relative to initial scores. This suggests AI models can enhance efficiency, reduce time of emotion lexicon development, and advance the field of emotion detection and sentiment analysis.

AUTHOR CONTRIBUTIONS

In the scope of this study, Senem KUMOVA METİN contributed to formation of the idea and the design, conducting of the analyses. Hande AKA UYMAZ contributed to formation of the design, collecting the data and conducting the analyses and literature review. The paper is written by both authors; all authors had approved the final version.

FUNDING

This work is carried under the grant of Izmir University of Economics - Coordinatorship of Scientific Research Projects, Project No: BAP2022-6, Building a Turkish Dataset for Emotion-Enriched Vector Space Models.

ACKNOWLEDGEMENTS

The authors wish to thank anonymous annotators for their great effort and time during the annotation process.

REFERENCES

Aka Uymaz, H., & Kumova Metin, S. (2022). Vector based sentiment and emotion analysis from text: A survey. In *Engineering Applications of Artificial Intelligence* (Vol. 113). Elsevier Ltd. <https://doi.org/10.1016/j.engappai.2022.104922>

Atapattu, T., Herath, M., Elvitigala, C., de Zoysa, P., Gunawardana, K., Thilakaratne, M., de Zoysa, K., & Falkner, K. (2022). *EmoMent: An Emotion Annotated*

- Mental Health Corpus from two South Asian Countries*. <http://arxiv.org/abs/2208.08486>
- Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*. <http://arxiv.org/abs/2303.04226>
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). *Evaluating Large Language Models Trained on Code*. <http://arxiv.org/abs/2107.03374>
- Day, M. Y., & Shaw, S. R. (2021). AI Customer Service System with Pre-trained Language and Response Ranking Models for University Admissions. *Proceedings - 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science, IRI 2021*. <https://doi.org/10.1109/IRI51335.2021.00062>
- Gala, N., & Brun, C. (2012). *Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions* (Vol. 2). TALN. <http://polarimots.lif.univ-mrs.fr>
- Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. <http://sentic.net/>
- Jacob Cohen. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 37–46.
- Joseph L. Fleiss. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Kertkeidkachorn, N., & Shirai, K. (n.d.). *Sentiment Analysis using the Relationship between Users and Products*. <https://github.com/knatthawut/gnnlm>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. <http://arxiv.org/abs/1907.11692>
- Mohammad, S. M., & Turney, P. D. (2010). *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. <http://www.wjh.harvard.edu/>
- Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a Word-Emotion Association Lexicon*. <http://arxiv.org/abs/1308.6297>
- OpenAI. (2023). *ChatGPT (Mar 14 version) [Large language model]*. <https://Chat.Openai.Com/Chat>
- Pérez, J., Díaz, J., García-Martin, J., & Tabuenca, B. (2020). Systematic literature reviews in software engineering—enhancement of the study selection process using Cohen's Kappa statistic. *Journal of Systems and Software*, 168. <https://doi.org/10.1016/j.jss.2020.110657>
- Pham, H. H., Nguyen, H. Q., Nguyen, H. T., Le, L. T., & Lam, K. (2023). *Evaluating the impact of an explainable machine learning system on the interobserver agreement in chest radiograph interpretation*. <http://arxiv.org/abs/2304.01220>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. <https://github.com/openai/gpt-2>
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). *Zero-Shot Text-to-Image Generation*. <http://arxiv.org/abs/2102.12092>
- Robert Plutchik. (1980). A general psychoevolutionary theory of emotion. *Plutchik, R., Kellerman, H. (Eds.), Theories of Emotion*. Academic Press, 3–33.
- Schuff, H., Barnes, J., Mohme, J., Padó, S., & Klinger, R. (2017). *Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus*. <http://www.ims.uni-stuttgart.de/data/>
- Segura Navarrete, A., Martínez-Araneda, C., Vidal-Castro, C., & Rubio-Manzano, C. (2021). A novel approach to the creation of a labelling lexicon for improving emotion analysis in text. *Electronic Library*, 39(1), 118–136. <https://doi.org/10.1108/EL-04-2020-0110>
- Staiano, J., & Guerini, M. (2014). *DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News*. Association for Computational Linguistics. <http://nie.mn/QuD17Z>
- Strapparava, C., & Valitutti, A. (2004). *WordNet-Affect: an Affective Extension of WordNet*. <https://www.researchgate.net/publication/254746105>
- Tocoglu, M. A., & Alpkocak, A. (2018). TREMO: A dataset for emotion analysis in Turkish. *Journal of Information Science*, 44(6), 848–860.
- Toçoglu, M. A., & Alpkocak, A. (2019). Lexicon-based emotion analysis in Turkish. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(2), 1213–1227. <https://doi.org/10.3906/elk-1807-41>
- Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: Definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7. <https://doi.org/10.1186/1471-2105-7-356>