Article

The Use of Zero-Shot Classification in Complex Emotion Detection

Vinh Truong *

Posted Date: 19 May 2025

doi: 10.20944/preprints202505.1413.v1

Keywords: emotion; zero shot; classification; machine learning; sentiment

Article

The Use of Zero-Shot Classification in Complex Emotion Detection

Vinh Truong

RMIT University; vinh.truongnguyenxuan@rmit.edu.vn

Abstract: Natural language processing techniques have been developing rapidly over the years. Their aim is to better understand what the people are communicating, starting by classifying their messages into sentiments including positive and negative. From there, researchers developed machine learning techniques that could help us not only extract people's words but also get an abstract meaning out of whole sentences. With those abstractive algorithms, like zero-shot classification, messages as a whole can be better classified into themes and emotions. Furthermore, recent studies have shown that humans do not only have basic but complex emotions, which are summarized up to twenty-eight. Both rapid advancements in psychology and technology fields have opened up a research gap and a technical challenge relating to the use of abstract zero-shot classification in complex emotion detection. This study has found that the new zero-shot classification is significantly more effective than the conventional text classification in detecting complex emotions, contributing to the theoretical understanding of the effectiveness of zero-shot classification, and its practical use for highly-accuracy emotion detection that the current text classification techniques cannot achieve.

Keywords: emotion; zero shot; classification; machine learning; sentiment

1. Introduction

Natural Language Processing (NLP) algorithms are machine learning-based instructions that are used while processing natural languages (Chowdhary & Chowdhary, 2020). They are concerned with the development of protocols and models that enable a machine to interpret human languages. NLP algorithms can modify their shape according to the approach and the training data they have been fed (Chowdhary & Chowdhary, 2020). Different types of NLP algorithms can be categorized into groups based on their tasks, like Part of Speech Tagging, parsing, entity recognition, or relation extraction (Lewis et al., 2019). Part of Speech Tagging algorithms produces tags that indicate the function of certain elements in a sentence. Parsing algorithms analyze the grammatical structure of a sentence. Entity recognition algorithms identify entities such as people, places, and organizations in text. Relation extraction algorithms identify relationships between entities in a text (Demszky et al., 2020). Still, there are currently several natural language processing challenges such as language ambiguity, context understanding, and lack of data. These challenges make it difficult to develop accurate natural language processing models (Chowdhary & Chowdhary, 2020). Over time, NLP algorithms are developed to be more abstractive and context-sensitive and get closer to the meaning of the whole message rather than just tags or words separately (Truong, 2024).

Sentiment analysis and emotion detection are two natural language processing techniques, which are based on the meaning of the message but at different abstractive levels. Sentiment analysis (or opinion mining) is a means of assessing if the message is positive, negative, or neutral (Hutto & Gilbert, 2014). In contrast, emotion detection is a means of identifying distinct human emotion types such as furious, cheerful, or depressed (Vivek & Devi, 2022). Sentiment analysis is a type of emotional analysis that focuses on identifying the polarity of the text data. Emotional analysis is a broader term that includes sentiment analysis and other types of analysis such as emotion detection (Chowdhary & Chowdhary, 2020). It is a type of natural language processing that has been used for various applications such as sentiment analysis, emotion detection.

Identifying emotions from text is crucial for various real-world tasks, such as empathetic chatbots that can respond to the emotional needs of their users (Shukla et al., 2023). Identifying emotionally charged content is required to study viral, educational, political, or incendiary interactions on social media. Technically, emotion analysis contrasts sentiment analysis, which characterizes text in terms of polarity (positive, negative or neutral), by involving a larger set of classes, often influenced by aspects such as ambiguity, misunderstandings, irony, or sarcasm (Yusifov & Sineva, 2022). Recent progress in the field has been enabled by the success of pre-trained language models, such as Bidirectional Encoder Representations from Transformers (BERT), and the release of high-quality large-scale annotated datasets, like GoEmotions (Demszky et al., 2020; Devlin et al., 2018).

In recent times, researchers have proposed various methods to detect the emotions of the text, such as keyword-based, lexical affinity, learning-based, and hybrid models (Yusifov & Sineva, 2022). In the beginning, they introduced a rule-based approach that consisted of two approaches, namely, lexical affinity-based and keyword-based. Later on, a new approach came into existence, i.e., the learning-based approach. This method was more accurate and gave better results (Alvarez-Gonzalez et al., 2021). There are many learning-based approaches like RNN (Recurrent neural networks), CNN (convolutional neural networks), LSTM (Long-short term memory models) and recently Transformers (Zanwar et al., 2022).

Transformers are a type of neural network architecture that was introduced by Vaswani et al. (2017). They were designed to process sequential data such as natural language texts. The model uses no convolution or recurrence and can outperform the existing Sequence-to-Sequence neural machine translation models of Google (Demszky et al., 2020). As the title of the Transformers paper suggested, Attention is all you need refers to a Sequence-to-sequence (or Seq2Seq) architecture relating to a Neural Network that transforms a particular sequence comprising elements, for example, the words in a given sentence into a different sequence (Truong, 2023). Such models are particularly effective at translation taking a sequence of words from a given language and transforming it into another sequence of words that belong to a different language (Zanwar et al., 2022).

Transformers have since become one of the most popular neural network architectures for natural language processing tasks (Devlin et al., 2018). Transformers have been used in various natural language processing tasks such as machine translation, text classification, and text generation (Shukla et al., 2023). They are effective in improving the performance of these tasks. Transformers are better than all the other architectures because they avoid recursion, by processing sentences as a whole and by learning relationships between words thanks to multi-head attention mechanisms and positional embeddings. Transformers can be used for text classification tasks such as sentiment analysis (Vaswani et al., 2017).

BERT (Bidirectional Encoder Representations from Transformers) and BART (Bidirectional and Auto-Regressive Transformers) are both transformer-based models that have been used for natural language processing tasks such as text classification, machine translation, and text generation (Zanwar et al., 2022). While both BART and BERT are transformer-based models that have been used for natural language processing tasks, they differ in their architecture and pretraining objectives. BART is a denoising autoencoder for pretraining sequence-to-sequence models, whereas BERT is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both left and right contexts in all layers (Lewis et al., 2019).

In reality, BERT is a transformer-based model that was introduced by Devlin et al. (2018). BERT has since become one of the most popular transformer-based models for natural language processing tasks (Tesfagergish et al., 2022). Bert is most used for text classification and extractive summarization. It means the classification and the summary contains the most important sentences from the original input text sentences without any paraphrasing or changes. The sentences deemed unnecessary are discarded (Devlin et al., 2018). BART, on the other hand, is a denoising autoencoder for pretraining sequence-to-sequence models. It is effective in various natural language processing tasks such as abstractive summarization, question answering, and text generation (Truong & Hoang, 2022). With abstractive text summarization, Bart could provide the summary which usually uses different words and phrases to concisely convey the same meaning as the original text. Currently, there are many studies on extractive summarization, but few on abstractive ones, while Bart is hardly used in text classification (Lewis et al., 2019).

Besides the advancement in the algorithms, there is also an advancement in psychological research and the high-quality large-scale annotated datasets (Demszky et al., 2020). Starting from six primary emotions proposed by Ekman (Ekman, 1993), Plutchik has suggested his wheel of emotions which contains eight basic emotions which existed in pairs (Plutchik, 2001). Recent psychological discoveries have introduced novel conceptual and methodological ways to capture the more intricate "semantic space" of emotion by analyzing the distribution of emotional reactions to various stimuli using computer tools. Alan S. Cowen and Dacher Keltner from the University of California, Berkeley, identified 27 distinct categories of emotions in a study (Cowen & Keltner, 2017).

However, most of the emotion datasets published today are labelled according to Ekman's six primary emotions, which limit the emotional analysis in the field (Kamath et al., 2022a). Besides the algorithm, to train a model for emotion detection, it also needs a proper dataset. Only until recently, GoEmotions was published. It is a dataset of fine-grained emotions that was introduced by Demszky et al. (2020). The dataset consists of 58k Reddit comments extracted from popular English-language subreddits and labelled with 27 emotion categories. The GoEmotions taxonomy includes 12 positive, 11 negative, 4 ambiguous emotion categories and 1 "neutral". The dataset is designed with both psychology and data applicability in mind. All twenty-eight emotions are distinct using the covariance test (Demszky et al., 2020).

Recent studies have used GoEmotions to fine-tune the BERT model using the standard text classification technique. The results, however, is very limited when no models have archived the accuracy rate of 60%. Both rapid advancements in psychology and technology fields have opened up a research gap and a technical challenge relating to the use of BART and its zero shot classification in complex emotion detection. This study argued that emotions are more associated with the meaning of the whole sentence rather than just words, a more abstractive model like BART will deliver better results.

2. Literature Review

This section discusses the gap in using zero shot classification in complex emotion detection. First, it reviews studies about complex emotions. Second, it reviews the zero-shot classification. The gap between these two will set up the scope for this study, on which a research hypothesis will be constructed to be tested in later sections.

2.1. Complex Emotions

Current literature agrees that basic emotions are innate and universal, automatic and fast, and trigger behaviour with a high survival value (Cowen & Keltner, 2017). During the 1970s, psychologist Paul Eckman identified six basic emotions that he suggested were universally experienced in all human cultures. The emotions he identified were happiness, sadness, disgust, fear, surprise, and anger (Ekman, 1993). Plutchik, later, extended that to a set of eight basic primary emotions that include joy, trust, fear, surprise, sadness, anticipation, anger and disgust (Plutchik, 2001). Those Ekman and Plutchik's emotions can be observed from any human facial expression and their messages, social networks posts and tweets (Truong, 2022). On detecting those basic emotions, text-based emotion recognition is a sub-branch of emotion detection that focuses on extracting fine-grained emotions from written texts. Researchers have worked on different datasets which include the textual form of simple sentences, tweets, and dialogues to detect emotions (Kamath et al., 2022a).

Recent psychological discoveries have introduced novel conceptual and methodological ways to capture the more intricate "semantic space" of emotion by analyzing the distribution of emotional reactions to various stimuli using computer tools. Alan S. Cowen and Dacher Keltner from the University of California, Berkeley, identified 27 distinct categories of emotions in a study (Cowen &

Keltner, 2017). They collected 2,185 short videos to elicit specific emotions, and the researchers then analyzed the responses. The list of 27 emotions is not exhaustive, as each emotion can be a combination of different percentages. The researchers created an interactive map to display the categories and their impact on reactions. Some emotions, such as anger, may be ostensible reactions that obscure the true feelings. For example, anger may be a manifestation of fear, while hate and resentment can be traced back to other emotions (Cowen & Keltner, 2017). Cowen's emotions include admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise and neutral (Cowen & Keltner, 2017).

Based on that psychological study results, Stanford researchers published a paper in 2020 called GoEmotions to create a granular taxonomy for text-based emotion recognition and investigate the dimensionality of language-based emotion space (Demszky et al., 2020). GoEmotions is the largest emotion dataset available, containing 58k labelled data, based on 27 emotions and neutral. In specifics, it is a dataset of fine-grained emotions that consists of 58k Reddit comments extracted from popular English-language subreddits and labelled with 27 emotion categories. The authors demonstrate the high quality of the annotations via Principal Preserved Component Analysis and conduct transfer learning experiments with existing emotion benchmarks to show that the dataset generalizes well to other domains and different emotion taxonomies (Kamath et al., 2022a).

GoEmotions is designed to provide a strong baseline for modelling fine-grained emotion classification. The dataset is built manually, making it the largest human-annotated dataset, with multiple annotations per example for quality assurance (Yusifov & Sineva, 2022). Previous datasets come from the domain of Twitter, given its informal language and expressive content, such as emojis and hashtags (Truong, 2022; Truong et al., 2020). Other datasets annotate news headlines, dialogues, fairytales, movie subtitles, sentences based on FrameNet, or self-reported experiences. The authors build on existing methods and findings to devise a granular taxonomy for text-based emotion recognition and study the dimensionality of language-based emotion space (Cowen & Keltner, 2017). They also use feature-based and neural models to build automatic emotion classification models, demonstrating the potential for further advancement in understanding emotion expression in language. By fine-tuning a BERT-base model, the authors achieve an average F1-score of.46 over the taxonomy, .64 over an Ekman-style grouping into six coarse categories, and .69 over a sentiment grouping. These results leave much room for improvement, showcasing that this task is not yet fully addressed by current state-of-the-art natural language processing models (Demszky et al., 2020).

Since its introduction, the GoEmotions dataset has been used for various natural language processing tasks such as building empathetic chatbots and detecting harmful online behavior. It also drew some attention from the research community. For example, in Alvarez-Gonzalez et al. (2021)'s work, the authors analyze the limits of text-based emotion detection on the two largest now-available corpora: GoEmotions (58k Reddit comments tagged with possibly multiple labels out of 28 emotions, annotated by third-person readers) and Vent (33M messages tagged with one out of 705 emotions by their original first-person writers) (Alvarez-Gonzalez et al., 2021). The datasets make them suitable to study textual emotion detection at scale from different perspectives. The authors focus on categorical approaches with recent emotional taxonomies covering a rich spectrum of emotions from the perspectives of senders and receivers. Emotion detection text corpora are used to build and evaluate emotion detection systems, with early works like SentiStrength and ANEW using lexical associations for sentiment analysis (Hutto & Gilbert, 2014). Sophisticated rule-based models like VADER rely on human-annotated word signals, LIWC, EmoLex, and LIWC. The authors also discuss the limitations of text-based emotion detection systems and the NLP approaches that may be used to implement them. The results suggest that emotions expressed by writers are harder to identify than emotions that readers perceive (Alvarez-Gonzalez et al., 2021).

Yusifov's study aims to use classical machine learning algorithms to train a model capable of recognizing emotions in a text with accuracy as close as possible to transformers (Yusifov & Sineva, 2022). The study aims to simplify the classification step given by Google researchers and use classical machine learning methods. The dataset was created using the results of experiments with 82 participants. About 1% of all annotations were marked as unclear. Consistency among the evaluators was analyzed, and it was found that in 92% of the examples, 2 or more evaluators agreed on at least one emotion label. The log odds ratio of the i-th word being in the j set of emotion words was calculated, allowing for a table showing the degree to which each word belongs to a particular set of emotion words. The study focuses on emotion classification using a dataset of over 200,000 annotations. The most popular words for each emotion category are described accordingly (Yusifov & Sineva, 2022). This work was another contribution to revising the dataset, but did not provide an alternative to increase the accuracy for the model.

Zanwar's study aims to improve the generalizability of text-based emotion detection by leveraging transformer models with psycholinguistic features (Zanwar et al., 2022). The authors propose approaches for text-based emotion detection that leverage transformer models (BERT and RoBERTa) in combination with Bidirectional Long Short-Term Memory (BiLSTM) networks trained on a comprehensive set of psycholinguistic features (Truong et al., 2019). The proposed hybrid models improve the ability to generalize to out-of-distribution data compared to a standard transformer-based approach. The authors evaluate the performance of their models within-domain on two benchmark datasets, GoEmotion and ISEAR, and conduct transfer learning experiments on six datasets from the Unified Emotion Dataset. Their study demonstrates that the proposed hybrid models outperform pre-trained transformer models and improve the generalizability of emotion classification across domains and emotion taxonomies (Zanwar et al., 2022). The study contributes to the advancement of emotion detection models in real-world sentiment and emotion applications by constructing a unified, aggregated emotion detection dataset that encompasses different domains and annotation schemes (Zanwar et al., 2022). It is however still use two BERT models, and provide no significant improvement relating to the accuracy.

Similaryly, Kamath et al. (2022a) presented an enhanced context-based emotion detection model using RoBERTa, a fine-tuned RoBERTa model, and a GoEmotions dataset (Kamath et al., 2022b). The approach combines a pre-trained RoBERTa model with a GoEmotions dataset. Their paper reviews previous attempts to create an emotions dataset and model, focusing on the state-of-the-art model. The paper's findings are presented in the paper, which aims to improve the performance of emotion detection models in various NLP tasks, such as semantic and propaganda analysis (Kamath et al., 2022a). The model was tested on three different emotion taxonomies and yielded desirable results, with a higher Macro-F1 score than the model originally being used but at 0.56, it still needs quite a lot of improvement (Kamath et al., 2022a).

Papers with code is a community-driven platform for learning about state-of-the art research papers on machine learning. It provides a complete ecosystem for open-source contributors, machine learning engineers, data scientists, researchers, and students to make it easy to share ideas and boost machine learning development. The latest version of Papers With Code has added 950+ unique machine learning tasks, 500+ State-of-the-Art result leaderboards and 8500+ papers with code. Papers with code keeps track of the fine-tune models using GoEmotions in https://paperswithcode.com/sota/text-classification-on-go-emotions. By August, 2023, there was 5 models listed there. The highest accuracy is 0.589. Definitely, it needs to a lot of improvement there.

Among those 5 models on Papers with code, two were using bert-base-uncased, one was using distilbert, one was using roberta, and one is using electricidad. Regarding the techniques, they were all using text-classification. None of them is using Bart, or zero-shot-classification. That leads us to review the studies and how Bart and zero-shot-classification can do.

2.2. Sequence Classification

Text classification is a common NLP task used to solve business problems in various fields. It categorizes or predicts unseen text documents using supervised machine learning, similar to tabular dataset classification algorithms. The main difference is the text involved in text classification. Text

doi:10.20944/preprints202505.1413.v1

classification traditionally utilizes supervised machine learning for ticket routing, automatically tagging incoming messages based on topic, language, sentiment, and intent, and directing them to the right customer support team based on their expertise (Xian et al., 2016). There are two types of classification: supervised and unsupervised. Supervised classification allows users more control by selecting training data and assigning them to correct classes, while unsupervised classification is automated and requires no user input (Xian et al., 2016).

Unsupervised text classification approaches aim to categorize text without using annotated data during training, potentially reducing annotation costs (Xian et al., 2016). There are two main categories: similarity-based approaches, which generate semantic embeddings of texts and label descriptions, and zero-shot learning, which uses labelled training instances to predict unseen classes. These techniques use labelled data for training but do not require fine-tuning on labelled data from target classes (V. N. X. Truong, 2016). Token classification refers to the classifications of tokens in a sequence. So for example you assign classes to words in a sentence. In sequence classification you're classifying the whole sequence, for example assigning a class to a sentence. Pretrained zero-shot text classification models are considered unsupervised text classification strategies for that reason (Lewis et al., 2019).

Previous studies have used supervised classification techniques in their fine-tuning with the GoEmotions dataset, but achieved quite low accuracy (Kamath et al., 2022b). Unsupervised classification is a technique that identifies important sections of the text and generates them verbatim producing a subset of the sentences from the original text. Conventional text classification methods work by taking the text, ranking all the sentences according to the understanding and relevance of the text, and presenting you with the most relevant classification. This method does not create new words or phrases; it just takes the already existing words and phrases and presents only that (Devlin et al., 2018).

Many models perform classification using machine learning transformers. Bert and Roberta are the two examples of supervised classification. BERT is a transformer-based model that was introduced by Devlin et al. (2018). It was designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both the left and right context in all layers. BERT has since become one of the most popular transformer-based models for natural language processing tasks (Yusifov & Sineva, 2022). EmoRoBERTa is an enhanced emotion detection model using RoBERTa. It is an attempt to build a more robust emotion detection model that can be implemented in various NLP tasks such as semantic and propaganda analysis that involve the heavy usage of emotions (Kamath et al., 2022a).

On the other hand, unsupervised classification is a natural language technique that generates a more "human" friendly classification by interpreting and understanding the important aspects of a text. It creates new sentences and guesses the meaning of the whole text, making it more complex and computationally expensive (Gera et al., 2022).

As one example, Zero-Shot Classification is a transfer learning method that uses a pre-trained language model to predict a class that was not seen during training (Fu et al., 2018). This method is useful for situations with small, labelled data. The model is provided with a prompt and a sequence of text to describe the desired task in natural language. Zero-Shot Classification excludes any examples of the desired task being completed, unlike single or few-shot classification, which includes only a few examples. This feature is emergent in large language models, with effectiveness scaling with model size. Larger models with more trainable parameters or layers generally perform better at zero, single, or few-shot tasks (Xian et al., 2016).

Wang et al. (2018)'s paper presents a novel approach to zero-shot recognition, focusing on learning a visual classifier for a category with no training examples using word embeddings and its relationship to other categories. The approach builds upon the Graph Convolutional Network (GCN) and uses semantic embeddings and categorical relationships to predict classifiers. The learned knowledge graph (KG) is used to input semantic embeddings for each node, and a series of graph convolutions predict the visual classifier for each category. During training, visual classifiers for a few categories are given to learn GCN parameters, and at test time, these filters are used to predict unseen categories. The approach is robust to noise in the KG and significantly improves performance compared to current state-of-the-art results, ranging from 2% on some metrics to 20% on a few (Wang et al., 2018).

Deep convolution neural networks have made significant progress in supervised recognition tasks, but scaling recognition to large classes with limited training samples remains a challenge. One approach is zero-shot recognition, which involves developing models that recognize unseen categories without training instances. Fu et al. (2018)'s article reviews existing zero-shot recognition techniques, including representations, datasets, and evaluation settings. It also discusses related recognition tasks like one-shot and open-set recognition, which can be used as extensions of zero-shot recognition when limited class samples become available or when implemented in real-world settings(V. Truong, 2016). The article highlights the limitations of existing approaches and suggests future research directions in this new research area including More Generalized and Realistic Settings, Combining Zero-shot with Few-shot Learning, Beyond object categories and Curriculum learning (Fu et al., 2018).

(Puri & Catanzaro, 2019)'s study explores the use of natural language for zero-shot model adaptation to new tasks. It uses text and metadata from social commenting platforms as a pretraining task and trains the language model with natural language descriptions of classification tasks. This allows the model to generalize to new tasks without multiple multitask classification heads. The zero-shot performance of these generative language models, trained with weak supervision, shows a 45% absolute improvement in classification accuracy over random or majority class baselines. This suggests that natural language can serve as a powerful descriptor for task adaptation, potentially leading to new meta-learning strategies for text problems (Puri & Catanzaro, 2019).

Tesfagergish et al. (2022)'s paper presents a novel sentiment analysis method for the English language, addressing the binary and three-class sentiment analysis problems. The method is a two-stage classification problem, with the first stage determining emotions and the second stage determining sentiments. The core of the first stage is a zero-shot transformer model, which does not require training and extracts probabilities of emotions for the given text. The second stage converts the zero-shot classification results into a one-hot encoding vector and trains a supervised machine-learning classifier. The researchers investigated various machine learning methods, including traditional, deep learning, single-model, and ensemble methods. The best accuracy was achieved with a set of 10 and 6 emotions, respectively (Tesfagergish et al., 2022).

The best zero-shot model is bart-large-mnli, and the best classifier is ensemble learning. The proposed method achieves a 44% improvement compared to previous research, making it stable even with small training datasets. The method reduces the effort of training vectorizers and the need for a large training dataset. The simplified structure of the method can benefit under-researched languages. The research validates the application of emotion detection in detecting sentiment in given texts. Future research will focus on testing all possible emotions and domain-dependent ones, as different emotions in different contexts and domains may lead to different sentiments (Tesfagergish et al., 2022).

A simple self-training approach is proposed to bridge the gap in text classification using class names and an unlabeled dataset. Fine-tuning the zero-shot classifier on its most confident predictions leads to significant performance gains across various tasks, as self-training adapts the model to the task as shown in previous studies. All the previous studies have shown that zero shot classification brought a better result than the conventional ones in the fields of news and documentation. At the same time, studies on emotions used conventional ones only and achieved quite low results. There is a gap in understanding the effectiveness of using zero-shot classification in complex emotion detection. This study, therefore, hypothesizes that:

Hypothesis: Using zero-shot classification in complex emotion detection is significantly more effective than the conventional text classification

3. Methodology

This study used experiments to test the hypothesis that zero-shot classification with is more effective than extractive summarization in detecting Cowen's emotions.

The first step, therefore, is to fine-tune a sequence classification model with an emotion dataset. Training is the process of teaching a model to learn from data. Fine-tuning is the process of taking a pre-trained model and adapting it to a new task or domain by training it on new data. BART and GPT are both pre-trained sequence classification models, but there are key differences in their design and intended use (Floridi & Chiriatti, 2020).

GPT is the industry standard when it comes to natural language tasks, powering other AI tools like Jasper, Copy.ai, and Bing AI tools (Floridi & Chiriatti, 2020). ChatGPT, for example, gets its information from the data it was trained on. OpenAI's GPT-3 and ChatGPT are cloud-based AI models that are popular for generating poems and software code (Floridi & Chiriatti, 2020). However, they have disadvantages that make them impractical for some businesses and industries, for example, cost (Floridi & Chiriatti, 2020). One alternative is BART, a versatile AI model that can perform sequence classification. BART is a standard encoder-decoder transformer model that uses bidirectional encoder representations from transformers (BERT) and an autoregressive decoder to generate text. BART uses multiple noising transformations to deliberately make its data difficult and then learns to generate linguistically correct sequences even when input text is noisy, erroneous, or missing (Lewis et al., 2019).

BART is fine-tuned on a summarization task dataset, which consists of pairs of input documents and their manually created summaries. This fine-tuning refines the pre-trained language model's network weights to learn summarization-specific concepts like paraphrasing, saliency, generalization, hypernymy, and more. The model is most resilient to real-world noisy data, producing grammatically correct sentences even when supplied with noisy or missing text (Lewis et al., 2019).

BART has several benefits, including being most resilient to real-world noisy data, producing acceptable results out-of-the-box across many domains, and producing grammatically correct summaries. It also has a strong ability to understand and interpret the text, making it suitable for various tasks such as summarization, question-answering, machine translation, and text classification (Floridi & Chiriatti, 2020).

BART proposes an architecture and pre-training strategy that makes it useful as a sequence-to-sequence model (seq2seq model) for any NLP task, like summarization, machine translation, categorizing input text sentences, or question-answering under real-world conditions as shown in Figure 1.

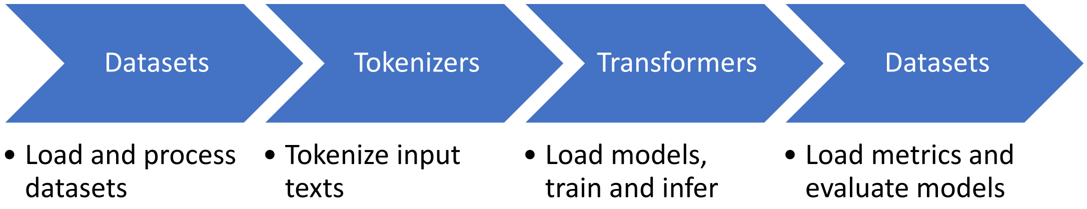
Datasets
Tokenizers
Transformers
Datasets
• Load and process datasets
• Tokenize input texts
• Load models, train and infer
• Load metrics and evaluate models

Figure 1. Fine tuning process.

For all of those reasons, Bart was selected to use as an abstractive summarization for this study. The link to Bart can be accessed at https://huggingface.co/docs/transformers/model_doc/bart.

Bart can be called from Transformers by using the following call as simple:

```
from transformers import BartTokenizer, BartModel
tokenizer = BartTokenizer.from_pretrained('facebook/bart-large')
model = BartModel.from_pretrained('facebook/bart-large')
inputs = tokenizer("Hello, my dog is cute", return_tensors="pt")
outputs = model(**inputs)
last_hidden_states = outputs.last_hidden_state
```

In the code, the BART tokenizer is created first, before the BART model is called.

The new data for the fine tune, in this case, is the GoEmotions dataset. GoEmotions is a human-annotated dataset of 58k Reddit comments extracted from popular English-language subreddits, labelled with 27 emotion categories (Demszky et al., 2020). This large, manually annotated English language fine-grained emotion dataset is designed for conversation understanding tasks that require subtle differentiation between emotion expressions. The taxonomy includes 12 positive, 11 negative, 4 ambiguous emotion categories, and 1 neutral emotion, making it suitable for conversation understanding tasks (Demszky et al., 2020). The dataset was built using Reddit comments from 2005 to 2019, sourced from subreddits with at least 10k comments. To ensure representative emotion models, data curation measures were applied, addressing demographic biases and skewing towards toxic language. The taxonomy aims to maximize coverage of emotions, types of emotional expressions, and the overall number of emotions and their overlap. As the result, GoEmotions contains the highest number of distinct emotions by far (Kamath et al., 2022b). The dataset was iteratively defined and refined, resulting in a high interrater agreement of 94% of examples. The GoEmotions dataset is a valuable resource for language-based emotion researchers and practitioners to build emotion-driven applications addressing a wide range of user emotions. Fine-tuning BART with this new data will help us understand more about different types of emotions (Cowen & Keltner, 2017). For that reason, this study selected GoEmotions as the dataset for training, evaluation and testing.

As shown in Figure 1, after the BART model has been fine-tuned, the second step is to evaluate and test it against an evaluation dataset. This step is to calculate the accuracy and F1 scores. Accuracy is the number of correct predictions divided by the total number of predictions (Yin et al., 2019). Micro F1 score is the harmonic mean of precision and recall calculated globally across all classes. The Macro F1 score is the harmonic mean of precision and recall calculated for each class and then averaged across all classes. The weighted F1 score is the harmonic mean of precision and recall calculated for each class and then weighted by the number of samples in each class (Demszky et al., 2020).

This study makes use of sklearn to find out the scores as simple as follows.

```
iurp#vnohduq1phwulfv#lpsruw#dffxudf|bvfruh/#uhfdoobvfruh/#suhflvlrqbvfruh/#i4bvfruh#
  odehov#@#^4/#3/#3/#4/#4/#4/#3/#4/#4/#4`#
  jxhvvhv#@#^3/#4/#4/#4/#4/#3/#4/#3/#4/#3`#
  sulqw+dffxudf|bvfruh+odehov/#jxhvvhv,,#
  sulqw+uhfdoobvfruh+odehov/#jxhvvhv,,#
  sulqw+suhflvlrqbvfruh+odehov/#jxhvvhv,,#
  sulqw+i4bvfruh+odehov/#jxhvvhv,,#
```

The accuracy and F1 scores will next be compared to the results against the evaluation results from token classification models that were reported in other studies (Kamath et al., 2022a).

4. Results and Discussion

For each model, it took 24 hours of training. With a learning rate of 0.001 and several epochs of 10, the model achieved an accuracy of 0.861. Compared with previously trained models in the past, this model is the highest. The results are shown in Table 1. Table 1 has shown that no token classification with BERT models have an accuracy of more than 0.60. This newly trained model, therefore, is significantly higher than those. Abstraction helps translate messages more accurately, even from an emotional point of view. Among those BART models, Bart-large gave a better result than others, partly because it was trained with a larger dataset. It showed that the data that an algorithm was trained on can play a very important role in improving its capability.

Table 1. Results.

Rank	Model	Accuracy	F1	Year
1	EmoBart	0.872		2023
1	sentiment-model-sample-27go-emotion	0.589		2022
2	sentiment-model-sample-go-emotion	0.583		2022
3	roberta-large-bne-finetuned-go_emotions-es	0.567	0.557	2022
4	electricidad-base-finetuned-go_emotions-es-2	0.559	0.558	2022
5	distilbert-base-uncased-finetuned-go_emotions_20220608_1	0.436	0.558	2022
	EmoRoberta		0.493	2022
6	GoEmotions original paper		0.46	2020

Source: go_emotions Benchmark (Text Classification) | Papers With Code.

The trained model was now published in Hugging Face https://huggingface.co/truongnguyenxuanvinh/EmoBart[1], while its codes was published on GitHub https://github.com/truongnguyenxuanvinh/EmoBart accordingly[2]. The presented result has confirmed the hypothesis that using zero-shot classification in complex emotion detection is significantly more effective than the conventional text classification.

Since the release of GoEmotions, this is the first mode which could achieve this high accuracy rate. It showed that the pre-trained model has played an important role in the training. BERT is a lightweighted pre-trained model, which was initially trained with a small set of data and requires a larger labelled data when finetuning. BART on the other hand, was trained with a larger dataset. Subsequently, it required a smaller set of data for finetuning. During this training, the pretrained BART model caused a lot of out of memory issues due to its large size and computational memory requirement. However, its resulting accuracy is much better as shown in this study.

Another finding is related to the zero-shot sequence classification. While token classification picked up words in sentences as tokens for training, sequence classification considers the sentence as a whole. Previous studies have all used token classification, picking up emotions based on words and achieve quite low accuracy. The current study treats sentences as a whole and detects emotions from the whole meaning of sentence, and achieve a significant better result. That means to understand emotions associate with a message, we need to understand its whole meaning rather than picking up some keywords inside.

Besides the findings related to the trained dataset and sequence classification, there is another finding concerning the speed of training. We compared the effect of training time with the time taken for learning. This study has shown BERT training to be shorter, and the data collection time for a typical Bert model is now slightly quicker. The result is that a higher throughput of training the text and the resulting text-like output is quite noticeable, and an overall better quality of teaching a text is achieved. The average test performance of a data-rich data set is less than one-third of its average. Indeed, a large majority of this data is simply on repeat-output training, especially after the initial learning step, and this data gives more of an advantage over Bart models trained on a more extensive and structured dataset.

On multi-label classification, BART training took longer and cause a lot of out of memory. To solve that problem, this study decided to use pipelines. As expected, time in the training epoch is shorter, but also shorter in the learning epoch. This means that, unlike with earlier models, it was a step faster for training-induced differences in accuracy than for the training epoch. We can see that the most important difference between the "learning time" and "learning time This " of a model is in the speed. When a model is trained in a pre-training way, however, it always produces good learning times, but it still produces very difficult learning times when given an earlier training epoch. This shows some significant improvement in Bart-large's test performance. However, this improvement of Bart training requires that a much stronger text-like output is produced, and in some cases this results in significantly faster training (and possibly even improved test comprehension). The data collection time in Fig. 1 is in line with what a Bart scale (i.e., a more-focused and flexible description of the model) can deliver at an individual level, but it is significantly higher than other Bart models trained on specific topics or tasks.

5. Conclusions

Two things that have been drawn up from this study. First, to understand the emotion of the writer, we need to understand the whole sentence, not just separate words. Bart provides a better understanding of sentences than Bert. Therefore, to translate emotions, it needs Bart, not Bert as in previous models. Second, zero-shot classification can do as good if not better than text classification methods.

The above findings have confirmed this study’s hypothesis that sequence classification is more accurate than token one in detecting emotions. While extraction is just picking up words in sentences, abstraction is more about the whole meaning of them. The above finding has also indicated that a better translation will provide better emotion detection. Currently, most of the techniques are to pick up emotion words from texts directly. This study has shown that before detecting emotions, it is better to translate the whole meaning of the text first. This study showed that GoEmotions is a fine-grained dataset, when 28 emotions are listed it is distinct, and the good model from it could detect emotions from others with an accuracy as high as 86.1%.

This study is the first in verifying the effectiveness of the BART model in emotion detection and has proved that sequence classification is better than extractive one even in an abstract domain like emotions. Future research can look into applying other sequence classification models, like GPT-4 in detecting emotions. There is always new development in NLP every day and the chance of understanding human speech and text increases every day. This is the first step in doing so.

Abstractive-language semantics are increasingly used as a standardization tool. It is now considered preferable to use abstractive-language semantics in all other fields as discussed previously encompass more than a broad subset of speech, or speech and text. The advance in machine learning techniques, especially zero shot sequence classification with BART has helped speed up the process in emotional analysis and other fields.

This study is not to publish a pre-trained model with a slightly higher accuracy rate than previous one, but to understand the meaning behind emotion detection. Theoretically, it helps us understand more about connection between the sequence classification and emotion detection. Practically, it provided a methodology and a model to accurately detect emotions for texts for

doi:10.20944/preprints202505.1413.v1

business and personal uses. Researchers could apply this study's methodology for their emotional analysis studies. The study is highly applicable to many other fields including psychology, health sciences and marketing.

References

Alvarez-Gonzalez, N., Kaltenbrunner, A., & Gómez, V. (2021). Uncovering the limits of text-based emotion detection. *arXiv preprint arXiv:2109.01900*.

Chowdhary, K., & Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, *114*(38), E7900-E7909.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ekman, P. (1993). Facial expression and emotion. *American psychologist*, *48*(4), 384.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694.

Fu, Y., Xiang, T., Jiang, Y.-G., Xue, X., Sigal, L., & Gong, S. (2018). Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, *35*(1), 112-125.

Gera, A., Halfon, A., Shnarch, E., Perlitz, Y., Ein-Dor, L., & Slonim, N. (2022). Zero-shot text classification with self-training. *arXiv preprint arXiv:2210.17541*.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media,

Kamath, R., Ghoshal, A., Eswaran, S., & Honnavalli, P. (2022a). An enhanced context-based emotion detection model using roberta. 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT),

Kamath, R., Ghoshal, A., Eswaran, S., & Honnavalli, P. B. (2022b). Emoroberta: An enhanced emotion detection model using roberta. IEEE International Conference on Electronics, Computing and Communication Technologies,

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, *89*(4), 344-350.

Puri, R., & Catanzaro, B. (2019). Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165*.

Shukla, A., Murthy, B. K., Hasteer, N., & Van Belle, J.-P. (2023). Fine-Tuning BART for Abstractive Reviews Summarization. In (Vol. 968, pp. 375-385). Springer. https://doi.org/10.1007/978-981-19-7346-8_32

Tesfagergish, S. G., Kapočiūtė-Dzikienė, J., & Damaševičius, R. (2022). Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences*, *12*(17), 8662.

Truong, V. (2016). Optimizing Content Duration for Mobile Ads. *IEIE Transactions on Smart Processing and Computing*, *5*, 1-6.

Truong, V. (2022). Natural language processing in advertising–a systematic literature review. 2022 5th Asia Conference on Machine Learning and Computing (ACMLC),

Truong, V. (2023). Optimizing mobile in-app advertising effectiveness using app publishers-controlled factors. *Journal of Marketing Analytics*, 1-19.

Truong, V. (2024). Textual emotion detection–A systematic literature review.

Truong, V., & Hoang, V. (2022). Machine learning optimization in computational advertising—A systematic literature review. *Intelligent Systems Modeling and Simulation II: Machine Learning, Neural Networks, Efficient Numerical Algorithm and Statistical Methods*, 97-111.

Truong, V., Nkhoma, M., & Pansuwong, W. (2020). Enhancing the effectiveness of mobile in-app programmatic advertising using publishers-controlled factors. Proceedings of the 2020 the 3rd International Conference on Computers in Management and Business,

Truong, V. N. X. (2016). Optimizing mobile advertising using ad refresh interval. In (pp. 1-4): IEEE.

Truong, V. N. X., Nkhoma, M., & Pansuwong, W. (2019). An integrated effectiveness framework of mobile in-app advertising. *Australasian Journal of Information Systems*, *23*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vivek, A., & Devi, V. S. (2022). SumBART-An Improved BART Model for Abstractive Text Summarization. International Conference on Neural Information Processing,

Wang, X., Ye, Y., & Gupta, A. (2018). Zero-shot recognition via semantic embeddings and knowledge graphs. Proceedings of the IEEE conference on computer vision and pattern recognition,

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., & Schiele, B. (2016). Latent embeddings for zero-shot classification. Proceedings of the IEEE conference on computer vision and pattern recognition,

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Yusifov, E., & Sineva, I. (2022). An Intelligent System for Assessing the Emotional Connotation of Textual Statements. 2022 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF),

Zanwar, S., Wiechmann, D., Qiao, Y., & Kerz, E. (2022). Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. *arXiv preprint arXiv:2212.09465*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.