IDEM: The IDioms with EMotions Dataset for Emotion Recognition

Alexander Prochnow[†], Johannes Bendler[†], Caroline Lange[†], Foivos Tzavellos[†], Bas Göritzer[†], Marijn ten Thij[†] and Riza Batista-Navarro[†,‡]
[†]Department of Advanced Computing Sciences, Maastricht University, The Netherlands
[‡]Department of Computer Science, The University of Manchester, UK
{a.prochnow, j.bendler, caro.lange, f.tzavellos, b.goritzer}@student.maastrichtuniversity.nl
{m.tenthij, r.batista}@maastrichtuniversity.nl

Abstract

Idiomatic expressions are used in everyday language and typically convey affect, i.e., emotion. However, very little work investigating the extent to which automated methods can recognise emotions expressed in idiom-containing text has been undertaken. This can be attributed to the lack of emotion-labelled datasets that support the development and evaluation of such methods. In this paper, we present the IDioms with EMotions (IDEM) dataset consisting of a total of 9685 idiom-containing sentences that were generated and labelled with any one of 36 emotion types, with the help of the GPT-4 generative language model. Human validation by two independent annotators showed that more than 51% of the generated sentences are ideal examples, with the annotators reaching an agreement rate of 62% measured in terms of Cohen's Kappa coefficient. To establish baseline performance on IDEM, various transformer-based emotion recognition approaches were implemented and evaluated. Results show that a RoBERTa model fine-tuned as a sequence classifier obtains a weighted F1-score of 58.73%, when the sequence provided as input specifies the idiom contained in a given sentence, together with its definition. Since this input configuration is based on the assumption that the idiom contained in the given sentence is already known, we also sought to assess the feasibility of automatically identifying the idioms contained in IDEM sentences. To this end, a hybrid idiom identification approach combining a rule-based method and a deep learning-based model was developed, whose performance on IDEM was determined to be 84.99% in terms of F1-score.

Keywords: Emotion recognition, Idiomatic expressions, Transformer models, Generative language models

1. Introduction

An idiomatic expression (or idiom) is a sequence of words whose meaning is non-compositional, i.e., not deducible from the meaning of its individual words (University of Oxford, 2022). Examples of idioms include *"butterflies in one's stomach"* which refers to feeling anxious rather than the presence of actual butterflies, and *"weak at the knees"* which refers to being overwhelmed by a strong emotion such as desire or fear, rather than lack of physical strength. According to Nunberg et al. (1994), idioms are typically used in language to express an affective stance, i.e., a sentiment or an emotion.

Sentiment analysis is a text classification task aimed at determining whether a given piece of text written in natural language contains positive, negative or neutral sentiment (Ligthart et al., 2021; Nandwani and Verma, 2021). A related task to sentiment analysis is emotion recognition (also referred to as emotion detection), whereby the emotion contained within text is identified according to fine-grained categories (e.g., happiness, anger, anxiety, sadness) rather than just the three broad sentiment types.

Despite the fact that idioms have been typically used to convey emotion and the well-known challenges that they pose to natural language processing (NLP) due to their non-compositionality and ambiguity (Zeng and Bhat, 2021), very little work has been undertaken to investigate the performance of emotion recognition methods when provided with input text that contains idioms. One of the primary reasons for this is the lack of datasets consisting of idiom-containing sentences that have been annotated with emotion labels to support the development and evaluation of emotion recognition models. To address this gap, we seek to address the following research questions in this work: (1) "How can a large-scale dataset of idiom-containing sentences with their corresponding emotion labels be developed, and can generative language models be exploited to make this task less burdensome for human annotators?" and (2) "How well do different types of state-of-the-art transformer-based models perform on the emotion recognition task, when evaluated based on idiom-containing sentences?"

To address the first research question, we employed a GPT-4 large language model (OpenAI, 2023) to automatically generate sentences that include frequently occurring idioms in the Sentiment Lexicon of IDiomatic Expressions (SLIDE) dataset (Jochim et al., 2018), together with the labels pertaining to the predominant emotion in each sentence. After validating the reliability of the generated data, we addressed the second research question by conducting experiments to evaluate the performance of the following transformer-based emotion recognition approaches on the above-

LREC-COLING 2024, pages 8569–8579
20-25 May, 2024.

mentioned dataset: (1) fine-tuning of BERT-based classification models; (2) fine-tuning of prompt-learning models; and (3) zero-shot prompting of a generative model.

For each of the above-mentioned approaches, we compared the performance obtained by the models in *idiom-ignorant* and *idiom-aware* data configurations. In the idiom-ignorant configuration, a model is not provided with any information pertaining to the idiom contained in a given sentence, whereas the idiom itself and its definition are supplied to the model in the idiom-aware configuration. As the latter configuration assumes that the idiom and its definition have been pre-identified, we also assess the feasibility of automatically identifying the idioms contained in our generated sentences, by developing a hybrid approach consisting of rule-based and deep learning-based idiom identification methods. The contributions of our work[1] include:

1. a novel dataset called IDEM (IDioms with EMotions) consisting of 9685 idiom-containing sentences, whereby each sentence is labelled with: (a) the idiom it contains, (b) the definition (meaning) of the idiom, and (c) the predominant emotion conveyed in the sentence, out of 36 emotion types;

2. an evaluation of different transformer-based emotion recognition approaches on IDEM, establishing baseline performance on the task of emotion recognition for idiom-containing sentences; and

3. an assessment of the feasibility of automatically identifying idioms contained within sentences, using a hybrid of a rule-based and a deep learning-based method.

In the remainder of this paper, we first review previously published research that is related to our work (Section 2). We then explain, in Section 3, how IDEM was constructed and how we assessed its reliability. In Section 4, we describe the emotion recognition approaches that we developed to establish baseline performance on the dataset. This is followed by a discussion of a rule-based and a hybrid approach that were implemented to assess the feasibility of automatically identifying idioms in the IDEM sentences. We then report and analyse the results of evaluating the emotion recognition models and idiom identification methods in Section 6. Finally, we summarise our findings and provide directions for future work in Section 7.

[1]Our dataset and code are publicly available at https://github.com/AlexanderProchnow/idem.

2. Related Work

As mentioned in the previous section, the lack of research into emotion recognition methods for idiom-containing sentences is largely due to the scarcity of resources that support the development and evaluation of such methods. For the purpose of assessing the extent to which idiom-based features (i.e., polarity) affect the performance of sentiment analysis models, Williams et al. (2015) developed the Idioment corpus which consists of 5980 sentences containing 580 idioms. However, the labels assigned to the sentences pertain to the sentiments they express, rather than finer-grained emotions. Meanwhile, the Sentiment Lexicon of IDiomatic Expressions (SLIDE) dataset (Jochim et al., 2018), developed by IBM, contains a much bigger collection of 5000 idioms, each of which is labelled with the sentiment associated with it as well as linked to its Wiktionary page. A shortcoming of this dataset, however, is that it is not accompanied by a corpus of idiom-containing sentences. Thus, it is unsuitable for training or evaluating models for analysing sentiments expressed within sentences. To the best of our knowledge, there exist no datasets that are comprised of idiom-containing sentences and the emotions that they convey; our work seeks to address this gap by providing a novel dataset of idiom-containing sentences labelled according to 36 emotion categories. Our decision to cover a much larger number of emotion types than the typical 6-10 emotion types considered by most of other datasets (Nandwani and Verma, 2021) was inspired by the work of Fokkinga and Desmet (2022) who proposed a rich typology of fine-grained emotions.

There exist a number of emotion recognition methods that are aimed at predicting the predominant emotions conveyed in text. Notably, the state-of-the-art methods for this task employ deep learning-based models. For instance, DialogueRNN, which was built upon recurrent neural networks (RNNs), was developed to detect emotions in conversations based on both textual features (from transcriptions of utterances) and audio-visual features (Majumder et al., 2019). Emotion recognition models that have emerged more recently are, however, based on the transformer architecture (Vaswani et al., 2017). For example, a Bidirectional Encoder Representations from Transformers (BERT) model was fine-tuned by Demszky et al. (2020) to train a multi-label classifier that can predict any number of emotions out of the 27 possible emotion types annotated in their GoEmotions dataset of Reddit posts. In another work, Alhuzali and Ananiadou (2021) proposed a neural architecture for span extraction whereby BERT was utilised to produce an embedding representation for a given text, which is then fed into a feed-forward network

that assigns an emotion category label to each token in the input text.

The emergence of large language models (LLMs) have enticed researchers to apply them to NLP tasks that previously required their own "narrow AI" model (i.e., a model trained specifically for one task). LLMs such as those from the Generative Pre-trained Transformer (GPT) family (Radford et al., 2018) have shown impressive performance on a variety of NLP tasks even in a zero-shot setting whereby a model is applied to a task without having been trained on task-specific training instances. To date, however, only the work of Venkatakrishnan et al. (2023) has explored their zero-shot application to emotion recognition, demonstrating that GPT-3.5 can identify nuances in the emotions expressed through language.

Although the methods that we present in this paper were developed without the intent to surpass state-of-the-art performance in emotion recognition but rather to establish baseline performance, they make for a novel contribution in that they demonstrate the extent to which different types of transformer-based models can identify emotions in sentences that contain idioms. As described in Section 4, apart from exploring approaches based on fine-tuning BERT-based classification models, we also investigated fine-tuning BERT-based models in a prompt-learning manner, and employing GPT-4 in a zero-shot setting.

3. Dataset Construction

In this section, we first provide an overview of the emotion annotation scheme that we have adopted, followed by a detailed description of how we exploited the GPT-4 large language model (LLM) to automatically generate idiom-containing sentences. Importantly, we explain how we assessed the reliability of the resulting dataset.

3.1. Emotion Categories

In contrast to most of the existing emotion-annotated datasets which were labelled according to 6-10 emotion types only (Plutchik, 1982; Ekman, 1992), we adopted a finer-grained emotion typology. This decision was informed by a study conducted by researchers at Delft University of Technology (TU Delft) which shows that a rich emotion typology forms the basis of emotion granularity, i.e., the ability to recognise nuances between emotions (Fokkinga and Desmet, 2022). We thus adopted their proposed typology which originally consists of 60 different emotions, 24 of which are positive emotions while the rest are negative ones. For every emotion in their typology, rich information is supplied, including the formal definition of the emotion. A visualisation of the emotion types is also provided, whereby emotions that are most similar to each other are shown in the same colour. Together with the provided definitions, the visualisation allowed us to observe that some of the emotion types overlap with each other (e.g., *insecurity* and *doubt*). By removing those overlapping emotion types (e.g., keeping *doubt* but not *insecurity*) and those that are very specific or rare (e.g., *schadenfreude*), we finally selected the 36 emotion types shown in Figure 1.

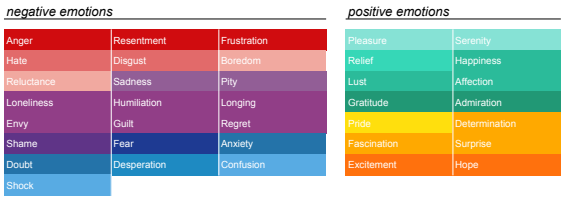

Figure 1: The set of 36 emotion types we adopted from the Emotion Typology developed at TU Delft (Fokkinga and Desmet, 2022).

Out of the 36 types in our emotion typology, 14 correspond to positive emotions while the remaining 22 pertain to negative ones. It is worth noting that in reducing the number of emotions from 60 to 36, we made an effort to ensure that the coarse-grained emotion types commonly used in the NLP research community (Nandwani and Verma, 2021) remain represented, as outlined in the mapping below, where the left-hand side is a course-grained emotion type and the right-hand side specify the corresponding types in our emotion typology.

- Anger: Anger, Resentment, Frustration, Hate, Disgust
- Boredom: Boredom, Reluctance
- Sadness: Sadness, Pity, Loneliness, Humiliation
- Desire: Longing, Envy
- Remorse: Guilt, Regret, Shame
- Fear: Fear, Anxiety, Doubt, Desperation, Confusion, Shock
- Joy: Pleasure, Serenity, Relief, Happiness, Lust, Affection, Gratitude, Admiration, Pride, Determination, Fascination, Surprise, Excitement, Hope

3.2. Generation and Labelling of Sentences

In order to create a sufficiently large dataset while minimising the human labour required and ensuring quality, we decided to employ a generative

Idiom	Definition (from Wiktionary)	Generated Sentence	Emotion
arm and a leg	Usually used after the verb 'cost', but also often 'charge', 'pay', and 'spend': a very high price for an item or service; an exorbitant price.	*It cost me an arm and a leg to repair my car after that accident.*	Frustration
go out of one's way	To make an extra effort, so as to help or hinder.	*He went out of his way to ignore me at the conference.*	Resentment
go bananas	To get angry; to go mad. To become silly or excited; to go crazy.	*When he saw the mess in his room, he went bananas.*	Anger

Table 1: Examples of sentences automatically generated and labelled with an emotion type by GPT-4.

language model, GPT-4, to automatically generate idiom-containing sentences and assign them emotion labels. The effectiveness of generating synthetic textual data using LLMs to support the development of NLP models has been previously demonstrated by Rosenbaum et al. (2022a,b) and Veselovsky et al. (2023).

Taking a subset of the idioms in the SLIDE dataset, we prompted GPT-4 to generate five sentences per idiom and assign an emotion label to each sentence, out of the 36 emotion types in our typology. The following template was populated to produce the prompt used as input to GPT-4: *You are good at generating sentences containing idioms and labelling them based on the emotion that they carry. The list of emotions is the following: <LIST OF 36 EMOTIONS>. Create 5 sentences for the idiom '<IDIOM>'. Label them according to the emotion.*

Table 1 presents a few examples from the set of generated sentences, together with the idiom they contain and the assigned emotion label. Additionally, we also supply the definition of the contained idiom as provided in its Wiktionary page (the link to which is available in the SLIDE dataset). In total, 11,610 sentences were automatically generated for 2322 idioms.

3.3. Data Reliability and Partitioning

To assess the reliability of the generated data, a randomly sampled subset of 1047 sentences (~10% of the total number of generated sentences) were manually validated by two annotators (Masters students with a good command of English) working independently. Each of them was asked to judge the quality of a generated instance (i.e., a sentence and the emotion label assigned to it) by categorising it as ideal or not. An instance is considered to be ideal only if the following conditions hold: (i) the sentence should include the idiom, (ii) the idiom is used in the sentence in an idiomatic sense (rather than its non-figurative meaning, where it exists), and (iii) the label assigned to the sentence should be one of our 36 emotion types, and should capture the emotion conveyed in the sentence.

Based on this validation process, it was found that 51% of the validated sentences (532 out of 1047) were considered by both annotators to be ideal. It is worth noting that the majority of the remaining 49% were considered to be non-ideal due to the presence of the emotion label in the sentences. These are not ideal as training or test examples since they explicitly mention the emotion; nevertheless, these sentences are not erroneous and thus have been included in our dataset. Measuring the agreement rate between the two annotators, we obtained a Cohen's Kappa value of 62%, which is considered to be substantial agreement (Landis and Koch, 1977).

As it became apparent during the annotation process that, in some cases, GPT-4 assigned a label that is not included in our set of 36 emotion types, we automatically discarded such sentences, leaving only 9685 out of the original full set of 11,160 sentences. This process also affected the number of sentences in the subset used for human validation, reducing it from 1047 to 956 sentences. We finally partitioned the full dataset into two subsets: a training set that consists of 8729 sentences, and a test set consisting of the 956 sentences that were manually validated.

4. Baseline Methods for Emotion Recognition

In this section, we describe each of the approaches that we implemented to establish the baseline performance of transformer-based emotion recognition models on IDEM. For each approach, two different data configurations were investigated: (1) idiom-ignorant (the input provided to the model does not include any information on the idiom), and (2) idiom-aware (the model is made aware of the idiom that is contained in a sentence, as well as its definition).

This allows us to investigate whether emotion recognition performance is affected by the inclusion of idiom-specific information in the input. Table 4 provides templates and examples illustrating how these data configurations were applied in each approach.

Approach	Idiom-ignorant	Idiom-aware
Fine-tuning of Sequence Classification Models	<SENTENCE>	<SENTENCE> This sentence includes the idiomatic expression '<IDIOM>'. The definition of this idiom is '<DEFINITION>'.
	He went out of his way to ignore me at the conference.	*He went out of his way to ignore me at the conference. This sentence includes the idiomatic expression 'go out of one's way'. The definition of this idiom is 'To make an extra effort, so as to help or hinder'.*
Fine-tuning of Sequence Pair Classification Models	<SENTENCE> [SEP] This sentence may or may not contain an idiomatic expression.	<SENTENCE> [SEP] The idiom is '<IDIOM>' which means '<DEFINITION>'.
	He went out of his way to ignore me at the conference. [SEP] This sentence may or may not contain an idiomatic expression.	*He went out of his way to ignore me at the conference. [SEP] The idiom is 'go out of one's way' which means 'To make an extra effort, so as to help or hinder'.*
Fine-tuning of Prompt-learning Models	<SENTENCE> The emotion of this sentence is ____	<SENTENCE> This sentence includes the idiomatic expression '<IDIOM>' which means '<DEFINITION>'. The emotion of this sentence is ____
	He went out of his way to ignore me at the conference. The emotion of this sentence is ____	*He went out of his way to ignore me at the conference. This sentence includes the idiomatic expression 'go out of one's way' which means 'To make an extra effort, so as to help or hinder'. The emotion of this sentence is ____*
Zero-shot Prompting of LLM	You identify the emotion expressed in a sentence and respond with one of <LIST OF 36 EMOTIONS>. The sentence is "<SENTENCE>"	You identify the emotion expressed in a sentence and respond with one of <LIST OF 36 EMOTIONS>. The sentence is "<SENTENCE>" This sentence contains the idiom '<IDIOM>'. The definition of this idiom is: '<DEFINITION>'.
	You identify the emotion expressed in a sentence and respond with one of <LIST OF 36 EMOTIONS>. The sentence is "He went out of his way to ignore me at the conference."	*You identify the emotion expressed in a sentence and respond with one of <LIST OF 36 EMOTIONS>. The sentence is "He went out of his way to ignore me at the conference." This sentence contains the idiom 'go out of one's way'. The definition of this idiom is: 'To make an extra effort, so as to help or hinder'.*

Table 2: Data configurations used in each approach. Templates are shown in white rows while examples are highlighted in grey.

4.1. Fine-tuning of Classification Models

Our first approach is based on fine-tuning transformer models for the downstream NLP task of multi-class classification, whereby a classification model is built upon a pre-trained transformer-based language model by placing a classification head on top of it. This model is then trained on instances that were labelled particularly for the classification task at hand, i.e., the emotion-labelled sentences in IDEM. We cast the problem in two ways, as outlined below.

Sequence Classification. In this classification task, the input is a sequence of tokens and the target output is the label that corresponds to the most predominant emotion conveyed in the sequence. Under the idiom-ignorant configuration where no idiom information is provided, the input sequence is simply an idiom-containing sentence, for example: *When he saw the mess in his room, he went bananas.* In contrast, to fine-tune models using the idiom-aware configuration that includes idiom information, we appended the idiom itself and its definition to the original input sequence, for example: *When he saw the mess in his room, he went bananas. This sentence includes the idiomatic expression 'go bananas'. The definition of this idiom is 'to get angry; to go mad. To become silly or excited; to go crazy.'*

Sequence Pair Classification. Different from the previously described classification task, sequence pair classification takes two separate sequences as input. A separator token is placed in between the two sequences, before they are presented to a model which then outputs an emotion label. In the idiom-ignorant configuration, the first sequence is the original idiom-containing sentence while the second sequence is: *This sentence may or may*

not contain an idiomatic expression. Meanwhile, under the idiom-aware configuration, the second sequence contains the idiom itself and its definition. An example input would thus be: *When he saw the mess in his room, he went bananas. [SEP] The idiom is 'go bananas' which means 'To get angry; to go mad. To become silly or excited; to go crazy.'*

We selected two transformer-based architectures to experiment with: BERT and RoBERTa (Liu et al., 2019). While BERT is the vanilla transformer architecture, RoBERTa is based on an improved and more effective pre-training procedure.

4.2. Fine-tuning of Prompt-learning Models

Prompt-learning models (PLMs) (Schick and Schütze, 2021; Liu et al., 2023; Zhou et al., 2023) facilitate the direct application of pre-trained transformer-based language models on downstream NLP problems without requiring the fine-tuning of a new task-specific model. This is achieved by reformulating the downstream problem (e.g., multi-class classification) as one of the original objectives learned during model pre-training, e.g., masked language modelling, with the use of a prompt. Instead of training a classification model, one would instead require a pre-trained transformer model to fill in the blank(s) in a prompt; the values provided by the model are then mapped to the target outputs, e.g., the class labels. Although it would have been possible to directly apply pre-trained transformer models in this zero-shot manner, we decided to fine-tune them by presenting training instances to the language models, in order to make the results of this approach comparable with those of fine-tuned classification models (described in Section 4.1). In the idiom-aware training configuration, an example prompt presented to the model is: *When he saw the mess in his room, he went bananas. This sentence includes the idiomatic expression 'go bananas' which means 'To get angry; to go mad. To become silly or excited; to go crazy'. Thus the emotion of this sentence is* ____. The prompt for the idiom-ignorant configuration is similar, except that it does not include the sentence specifying the idiom and its definition. The same transformer-based architectures that were employed in our first approach (fine-tuning classification models), were used as language models for prompt-learning: BERT and RoBERTa.

4.3. Zero-shot Prompting of LLM

As mentioned in Section 1, generative LLMs have shown impressive performance on a number of downstream NLP tasks. We thus employed an LLM, specifically GPT-4, in a zero-shot manner, whereby the model generates text as a response to a prompt (that might have never been encountered by the model before). Under the idiom-ignorant configuration, GPT-4 is given a prompt that specifies only the sentence, for example: *You identify the emotion expressed in a sentence and respond with one of ‹LIST OF 36 EMOTIONS›. The sentence is "When he saw the mess in his room, he went bananas."* Meanwhile, the prompt for the idiom-aware configuration includes both the idiom contained in the sentence and its definition, for example: *You identify the emotion expressed in a sentence and respond with one of ‹LIST OF 36 EMOTIONS›. The sentence is "When he saw the mess in his room, he went bananas." This sentence contains the idiom 'go bananas'. The definition of this idiom is: 'To get angry; to go mad. To become silly or excited; to go crazy.'*

4.4. Implementation

All of the above approaches were implemented using Python. In fine-tuning the classification models, the `simpletransformers` library[2] was used, which allowed for loading the pre-trained BERT and RoBERTa models, respectively `bert-base-cased`[3] and `roberta-base`[4], directly from Huggingface[5]. Each model was fine-tuned on the IDEM training set for 10 epochs, using a batch size of 16 for best processing efficiency and the default learning rate of 4e-5 with the AdamW optimiser. Meanwhile, the fine-tuning of prompt-learning models was facilitated by the OpenPrompt library[6], a prompt-learning framework that also provides direct access to the same pre-trained BERT and RoBERTa models. Here, the same hyperparameter values as above were adopted, except that 1e-4 was used as the learning rate, as it was found to result in optimal model training, based on initial, non-exhaustive experiments. Across all training runs, we took the model produced by the epoch where the best performance (in terms of error loss) on a held-out validation set (a subset of the training set) was obtained.

In implementing zero-shot prompting of GPT-4, we utilised the OpenAI Python library[7] which provides access to the OpenAI chat completions endpoint[8]. Default GPT-4 settings (e.g., temperature,

[2] https://github.com/ThilinaRajapakse/simpletransformers

[3] https://huggingface.co/bert-base-cased

[4] https://huggingface.co/roberta-base

[5] https://huggingface.co/

[6] https://github.com/thunlp/OpenPrompt

[7] https://github.com/openai/openai-python

[8] https://platform.openai.com/docs/api-reference/chat/create

Top P) were used. We conducted a non-exhaustive phase of prompt engineering, where the prompts were devised such that the GPT-4 outputs required little post-processing.

5. Automating Idiom Identification

All of the baseline emotion recognition methods that made use of the idiom-aware configuration (as described in the previous section), were based on the assumption that the idiom contained within a given sentence and its definition have been pre-identified. Thus, as an additional study, we sought to investigate the extent to which idioms in IDEM sentences can be automatically identified (and subsequently linked to their definition in Wiktionary). Given an idiom-containing sentence, an automated approach should be able to select the idiom that is used in the sentence, out of all the candidates in a dictionary of idioms, which, in our case, is the SLIDE dataset. This task can be challenging as idioms could appear in sentences in the form of different variations; for example, a different pronoun or verb tense could be used, or other words could be interspersed within the idiom. An example of a challenging case is the sentence *"He poured out his whole heart to a friend."*, whose tokens do not exactly match the idiom *"pour someone's heart out"*. To address this task, we developed a hybrid idiom identification approach that combines a rule-based and a deep learning-based method.

5.1. Rule-based Method

The first step involved in this approach is the lemmatisation of each of the idioms in the SLIDE dataset, as well as any given input sentence. An idiom is then considered to be a candidate match if all of its lemmatised tokens are present in the lemmatised sentence. To ensure that the lemmatised tokens do pertain to an idiom (rather than being present in the sentence only as a matter of coincidence), the following two scores were calculated.

Token gap score (t_g): the number of sentence tokens that appear in between the candidate idiom's tokens. This value is normalised by dividing it by the largest possible gap, i.e., if all other tokens are in between the candidate idiom's tokens. The normalised value is subtracted from 1 to obtain the final token gap score t_g. A t_g value that is closer to 1 means that the idiom's tokens appear closer to each other in the given sentence.

Token order score (t_o): the number of token bigrams that are common between the idiom and the given sentence divided by the total number of token bigrams in the idiom.

The two scores are then combined in the form of a weighted harmonic mean, F_β, calculated as:

$$F_\beta = (1 + \beta^2) \cdot \frac{t_g \cdot t_o}{(\beta^2 \cdot t_g) + t_o} \quad (1)$$

where the token order score t_o is given a slightly bigger emphasis by setting $\beta = 1.2$. If the harmonic mean is above 0.9, the candidate is considered to be an identified idiom.

5.2. Deep learning-based Model

After reviewing the literature, we decided to investigate the performance of the DISC (iDentifier of Idiomatic expressions via Semantic Compatibility) model (Zeng and Bhat, 2021), a state-of-the-art deep learning-based model for idiom identification. Designed as a Bi-LSTM-based sequence labelling model, DISC classifies every token in a given sentence as belonging to an idiom or not. It makes use of different types of embeddings to produce both the literal and contextual representations of a potential idiomatic expression. Part-of-speech tag embeddings are combined with GloVe embeddings (Pennington et al., 2014) to obtain the literal representation of a given sentence, while contextualised embeddings of the same sentence are generated using a BERT model. The semantic compatibility between these two representations is then determined by employing an attention mechanism, the result of which is used to finally identify the tokens that comprise an idiom. An advantage of the DISC model is that it does not rely on a predefined dictionary in order to identify idioms, hence it can potentially identify even idioms that are not catalogued in a lexicon like SLIDE.

Following the training procedure described in the original paper by Zeng and Bhat (2021) and using their original implementation[9], we trained our own DISC model on the MAGPIE dataset (Haagsma et al., 2020), utilising their training set that does not contain any idioms that appear in their test set[10].

5.3. Hybrid Approach

Our final idiom recognition model integrates the rule-based method and the DISC model described above. Firstly, for every given sentence, matching idioms are identified by applying the rule-based method, resulting in a set of candidate idioms. Separately, our trained DISC model is also applied on every given sentence. Since DISC casts idiom identification as a sequence labelling problem, its output is a subsequence of tokens extracted verbatim from

[9]Available at https://github.com/zzeng13/DISC

[10]Available at https://github.com/hslh/magpie-corpus

a given sentence. Thus, in cases where a variation of the idiom (as exemplified in Section 5) appears in the sentence, the extracted subsequence might not always exactly match the corresponding idiom in the SLIDE dataset. In order to identify the best-matching idiom, we post-processed the output of the DISC model by lemmatising it. The resulting lemmatised subsequence is then compared with the lemmatised form of every idiom in SLIDE; if an exact match is found, the matching idiom is considered to be a candidate idiom. Finally, we take the union of the candidates produced by both the rule-based method and the DISC model.

6. Results and Discussion

In this section, we present the results of evaluating our baseline emotion recognition approaches and the hybrid approach to idiom identification.

6.1. Emotion Recognition Evaluation

Table 3 presents the results of evaluating the various approaches to emotion recognition that we developed. Overall, GPT-4 obtained the best accuracy and weighted macro-averaged F1-score, i.e., 61.00% and 61.05%, respectively. It is, however, worth noting that this might very well be a result of the fact that the sentences comprising the IDEM test set were themselves generated by GPT-4, i.e., the same model.

Looking at the performance of our other models, one can observe that with respect to the approaches that are based on fine-tuned pre-trained language models (i.e., those that are not underpinned by GPT-4), RoBERTa consistently outperformed BERT, regardless of whether the idiom-ignorant or idiom-aware configuration was used. For example, under the idiom-aware configuration, RoBERTa obtained a weighted F1-score of 58.73%, which is 5.45 percentage points higher than BERT's F1-score of 53.28%. This is perhaps unsurprising as RoBERTa is based on an improved training procedure and has been shown to surpass the performance of BERT on many downstream NLP tasks (Casola et al., 2022). It is also noticable that while fine-tuned sequence classification and sequence pair classification models seem to obtain equally competitive peformance on the emotion recognition task, the performance of fine-tuned prompt-learning models is relatively poor. For instance, in the idiom-ignorant configuration, BERT and RoBERTa obtained weighted macro-averaged F1-scores of 48.67% and 52.46%, respectively, which are noticeably lower than the same models' performance when fine-tuned as sequence classifiers (53.55% and 57.85%) and sequence pair classifiers (53.09% and 58.52%). These results are consistent with some of the findings by Mosbach et al. (2023), whose work showed that fine-tuning of prompt-learning models led to poorer performance in comparison to the fine-tuning of task-specific models.

When comparing the weighted F1-scores obtained in the idiom-ignorant and idiom-aware configurations, no clear differences in performance can be seen. However, it is worth noting that, next to GPT-4, it is the RoBERTa-based sequence classification model that was fine-tuned using the idiom-aware configuration that obtained the best performance with an accuracy of 58.79% and weighted F1-score of 58.73%. Considering that RoBERTa is a much smaller model (125M parameters) than GPT-4 (1.76T parameters), and the fact that the latter is the same model that was used to generate the test data, it is impressive that the difference in their performance is only less than 3 percentage points.

6.2. Idiom Identification Evaluation

Table 4 presents the results of evaluating the rule-based and hybrid idiom identification approaches on the IDEM test set containing 956 sentences. As can be seen in the table, the rule-based method obtained satisfactory performance on the task, with an F1-score of 84.68%. Combining the results of the DISC model with those of the rule-based method via the hybrid approach led to marginal improvement, yielding an F1-score of 84.99%, with precision dropping slightly from 76.95% to 75.95%, but recall increasing from 94.13% to 96.47%. This demonstrates the feasibility of automatically identifying idioms in IDEM, even with just a rule-based method.

We sought to investigate the extent to which the use of automatically identified idioms affects the performance of emotion recognition models. Thus, we re-applied the RoBERTa-based sequence classification model that obtained an emotion recognition accuracy and F1-score of 58.79% and 58.73%, respectively (as discussed in Section 6.1) using the idiom-aware configuration, on the IDEM test set. This time, however, we provided it with the idioms automatically identified by our hybrid identification method (and their corresponding definitions) instead of pre-identified, gold standard idioms. Interestingly, the drop in performance is only minimal, decreasing by only about 1 percentage point: the accuracy and F1-score obtained are 57.85% and 57.67%, respectively. This implies that the hybrid identification method can form part of a fully automated pipeline for idiom-aware emotion recognition.

Approach	Model	Idiom-ignorant		Idiom-aware	
		Accuracy	**F1-score**	**Accuracy**	**F1-score**
Fine-tuning of Sequence Classification Models	bert-base-cased	54.18	53.55	54.50	53.28
	roberta-base	58.37	57.85	**58.79**	**58.73**
Fine-tuning of Sequence Pair Classification Models	bert-base-cased	53.87	53.09	52.62	51.84
	roberta-base	59.00	58.52	57.01	56.93
Fine-tuning of Prompt-learning Models	bert-base-cased	48.01	48.67	47.28	48.95
	roberta-base	52.20	52.46	50.73	52.20
Zero-shot Prompting of LLM	GPT-4	**61.00**	**61.05**	55.66	55.75

Table 3: Results of evaluating baseline approaches to emotion recognition on the IDEM test set, in terms of accuracy (%) and F1-scores (%). F1-scores are weighted over all classes, i.e., the 36 emotion types.

	Precision	**Recall**	**F1-score**
Rule-based	76.95	94.13	84.68
Hybrid	75.95	96.47	84.99

Table 4: Results of evaluating our idiom identification methods on the IDEM test set.

7. Conclusion and Future Work

This paper presents our work on developing IDEM, a new dataset consisting of idiom-containing sentences that are labelled with the predominant emotion that they convey. We report the following findings: (1) To reduce the manual effort typically required in building new datasets, GPT-4 can be employed to automatically generate and label idiom-containing sentences, although human validation is still necessary to remove hallucinations, i.e., sentences with labels that do not exist within our 36 emotion types. (2) Comparing different transformer-based baseline methods for emotion recognition, a RoBERTa model fine-tuned as a sequence classifier that is made aware of the idiom contained in a given sentence (and its definition) obtains competitive performance (an F1-score of 58.73%) relative to that of GPT-4 applied in a zero-shot manner (61.05%).

In our future work, we plan to enrich the labels in IDEM by allowing the assignment of multiple emotion types to each sentence, thus capturing cases where more than one emotion is conveyed in text. This, in turn, will support the development and evaluation of multi-label emotion recognition methods for idiom-containing sentences. We also encourage the NLP community to investigate the extent to which emotion recognition performance on IDEM can be improved with the use and comparison of, for example, other state-of-the-art model architectures and open-source LLMs.

Limitations

In this study, we focused only on English idioms given that English is the language for which idiom lexicons are most available. Idioms are highly language-dependent, and time and resource constraints did not allow us to build a dataset for other languages. It is also for reasons of time and resource constraints that we were able to employ only two dataset annotators.

Some bias might have been introduced in using GPT-4 in generating and labelling our dataset and then using the same model for zero-shot prompting. Although results of our experiments show that GPT-4 did not necessarily perform much better than our other baselines, especially in the idiom-aware configuration, one can explore other LLMs (e.g., Llama, Mistral) as part of a zero-shot prompting approach to emotion recognition, in order to mitigate the above-mentioned bias.

8. Bibliographical References

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Silvia Casola, Ivano Lauriola, and Alberto Lavelli. 2022. Pre-trained transformers: an empirical comparison. *Machine Learning with Applications*, 9:100334.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zulfadzli Drus and Haliyana Khalid. 2019. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Steven Fokkinga and Pieter Desmet. 2022. Emotion Typology. Available online: https://emotiontypology.com.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE—a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Languages Resources Association (ELRA).

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. 2021. Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, 54(7):4997–5053.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

OpenAI. 2023. GPT-4 Technical Report. Available online: https://arxiv.org/abs/2303.08774.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Marco Damonte, Isabel Groves, and Amir Saffari. 2022a. CLASP: Few-Shot Cross-Lingual Data Augmentation for Semantic Parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 444–462.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022b. LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

University of Oxford. 2022. Oxford Learner's Dictionaries. Available online: https://www.oxfordlearnersdictionaries.com/definition/english/idiom?q=idiom.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Radhakrishnan Venkatakrishnan, Mahsa Goodarzi, and M. Abdullah Canbaz. 2023. Exploring Large Language Models' Emotion Detection Abilities: Use Cases From the Middle East. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 241–244.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Yulin Zhou, Yiren Zhao, Ilia Shumailov, Robert Mullins, and Yarin Gal. 2023. Revisiting Automated Prompting: Are We Actually Doing Better? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1822–1832, Toronto, Canada. Association for Computational Linguistics.