UMUTeam at SemEval-2025 Task 1: Leveraging Multimodal and Large Language Model for Identifying and Ranking Idiomatic Expressions

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan, tomas.bernalb, joseantonio.garcia8, valencia}@um.es

Abstract

Idioms are non-compositional linguistic expressions whose meanings cannot be directly inferred from the individual words that compose them, posing significant challenges for natural language processing systems. This paper describes the participation of the UMUTeam in Subtask A of the AdMIRe shared task (SemEval 2025), which focuses on understanding idiomatic expressions through visual and contextual representations in English and Portuguese. Specifically, the task involves ranking a set of images according to how well they represent the sense of a potentially idiomatic nominal compound within a given contextual sentence. To address this challenge, we adopted a multimodal approach that combines textual and visual features using pre-trained language models, such as BERT and XLM-RoBERTa, along with Vision Transformers. Additionally, we explored the in-context learning capabilities of Large Language Models (LLMs), particularly Llama-3.1-8B, for image classification. These models are trained using a regression approach to rank images according to their semantic alignment with the contextual meaning of idioms. The results show that the Llama-3.1-8B model performs best for English, ranking 17th in the test set and 12th in the extended evaluation set, while the XLM + ViT model is more effective for Portuguese, ranking 9th in the test set and 8th in the extended evaluation set.

1 Introduction

An idiom is a linguistic expression or construction whose meaning cannot be derived directly and literally from the words that make it up (Bobrow and Bell, 1973). Idioms are fixed phrases or sayings that use a figurative sense to convey ideas, emotions, or situations in a particular language, and they can be difficult to translate into another language without losing their intended meaning. An extensive review of figurative language and idioms has been carried out in (del Pilar Salas-Zárate et al., 2020).

Idioms are common to all languages and are based on the culture, history and traditions of each society, reflecting the unique cultural and linguistic aspects of that society. This means that a literal interpretation of them results in the loss of essential cultural and contextual nuances, preventing their original meaning from being accurately conveyed in another language (Lakoff and Johnson, 1980). For example, the expression "when pigs fly" is an excellent example of non-literal language. Even when people are not explicitly familiar with the expression, human cognition is able to infer its meaning from the images it evokes. The impossibility of pigs flying intuitively suggests an event of extreme improbability or outright impossibility. This phenomenon highlights the sophisticated way in which humans process language. Rather than relying solely on literal definitions, humans integrate contextual cues, cultural knowledge, and metaphor to construct meaning.

Large Language Models (LLMs) have proven to be efficient in various natural language processing tasks such as hate speech detection (García-Díaz et al., 2023; Pan et al., 2024), emotion identification (Salmerón-Ríos et al., 2024), hope speech (García-Baena et al., 2023), or translation among others. However, they often misinterpret or mistranslate idiomatic expressions because they tend to process idioms as if they were compound expressions, producing literal translation errors, in which the figurative meaning is lost and the intended message is inappropriately conveyed (Li et al., 2024) (Tayyar Madabushi et al., 2021). Their reliance on such correlations, without a real conceptual understanding of non-compositional language, remains a fundamental limitation in dealing with idioms (Phelps et al., 2024).

Accurate interpretation of idioms is essential for applications such as sentiment analysis, ma-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 743–749
July 31 - August 1, 2025

chine translation and natural language understanding (Salehi et al., 2015) (Reddy et al., 2011). As these fields require a more sophisticated understanding of language, improving models' ability to interpret idioms could lead to significant improvements in their overall performance.

The AdMIRe shared task (SemEval 2025) (He et al., 2025) focuses on the discovery and understanding of idioms. However, this task is not about binary classification of idioms, but about understanding and correctly associating their meaning through visual and visual-temporal representations, in two languages: English and Portuguese. It is divided into two subtasks: (1) **Subtask A: Static Images**. Given a set of 5 images and a context sentence in which a given potentially idiomatic nominal compound (NC) appears, the goal is to rank the images according to how well they represent the sense in which the NC is used in the given context sentence; and (2) **Subtask B: Next Image Prediction**. Given a target expression and an image sequence from which the last of 3 images has been removed, the goal is to select the best fill from a set of candidate images.

Vision transformers have proven effective in correctly identifying and understanding objects, patterns, and contexts in complex images, facilitating advances in classification, detection, and segmentation in computer vision applications (Lu et al., 2019). This makes them efficient in the task of correctly identifying, understanding and graphically representing idioms present in an image. Moreover, thanks to the combined use of these models with a large language model, it is possible to identify the idioms present in the text while describing the images, which allows to find out which image best represents each idiom. Therefore, in this study, two different approaches have been tested to discover and understand idiomatic expressions using visual and visual-temporal representations in two languages, English and Portuguese. The first approach involves a multimodal model that uses the correlation between images and textual descriptions to capture the contextual meaning of idiomatic expressions. The second approach exploits the in-context learning capability of LLMs, such as Llama-3.1-8B (Dubey et al., 2024), to classify images based on their descriptions using prompt engineering techniques such as zero-shot learning.

We participated in Subtask A and used a multimodal approach combining textual (image caption) and visual representations. Based on pre-trained language models such as BERT and XLM-RoBERTa, and vision models such as Vision Transformers (ViT), we have adopted an approach that fuses textual embeddings and images to improve the semantic understanding of idioms in context. In addition, the models have been trained using a regression approach with the aim of learning to assign scores to images according to their correspondence with the contextual meaning of an idiom in a sentence. In addition, we tested the in-context learning capability of LLMs, such as Llama-3.1-8B, for image classification by utilizing image descriptions.

2 Background

At the computational level, one of the first approaches to deal with idiomatic expressions was the use of supervised binary classification models to distinguish between literal and figurative uses, as seen in the work of (Fazly et al., 2009), who proposed an unsupervised method to identify idiomatic types and occurrences from syntactic patterns and lexical co-occurrences.

Subsequently, with the rise of models of distributional representations, research such as (Salehi et al., 2015) explored the use of word embeddings to predict the compositionality of complex expressions. While useful, these approaches do not adequately capture deep idiomatic meaning, especially in ambiguous contexts.

More recently, with the development of pre-trained language models such as BERT, RoBERTa, T5, among others, considerable improvement has been achieved in detecting idioms using context. However, these models still face difficulties when attempting to represent the full semantic meaning of idioms, especially outside their immediate context.

Moreover, in terms of evaluation, tasks such as SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al., 2022) have proposed multilingual benchmarks. However, as highlighted by (Boisson et al., 2023), certain artifacts in these datasets may bias the results, allowing models to detect idiomatic expressions without true semantic understanding.

In the face of these limitations, there has been a growing interest in multimodal approaches that integrate text with images or visual descriptions. To this end, the AdMIRe task has emerged, which seeks to assess idiomatic comprehension through

```
**Instructions for the model:**
1. Identify the nominal compound (NC) in the provided context sentence.
2. Interpret the specific meaning of the NC **within the given context**
(considering both literal and idiomatic meanings).
3. Analyze the five provided image captions.
4. Compare each image caption against the meaning of the NC in context,
assessing **how well it represents the intended sense**.
5. Assign a **similarity score (1.000 - 10.000)** for each image, where:
   - **1.000** means the image is completely unrelated.
   - **10.000** means the image perfectly represents the meaning.
6. Output the similarity scores in the following format:
   ```
   Image_1: 3.245, Image_2: 7.530, Image_3: 9.870, Image_4: 5.610, Image_5: 2.490
   ```
7. Do **not** provide any explanations for the scores.
8. Ensure consistency in scoring across different instances.

Context: {sentence}
NC: {compound}
Image_1 caption: {image_caption_1}
Image_2 caption: {image_caption_2}
Image_3 caption: {image_caption_3}
Image_4 caption: {image_caption_4}
Image_5 caption: {image_caption_5}

**Output the similarity scores in the following format:**
```
Image_1: 3.245, Image_2: 7.530, Image_3: 9.870, Image_4: 5.610, Image_5: 2.490
```

Answer:
```

Listing 1: Structure of the prompt

multimodal representations in English and Portuguese.

3 System overview

Figure 1 shows the architecture of the first approach, which consists of training a multimodal model that combines visual and textual features to understand idiomatic expressions. Unlike classifying images in a binary problem using an input sentence (e.g., as literal or figurative), the goal of this approach is to correctly associate their meaning by using images and contextual textual descriptions.

The process begins with preprocessing the training set. In this case, we have linearly ranked the 5 images associated with each `sentence`, assigning them a relevance score based on their position in the expected order. This score is calculated in a normalized way, where the most relevant image receives the highest value and the least relevant the lowest. The ranking is based on the `expected_order` list, which indicates the expected sequence of relevance of the images for each idiomatic expression. For example, the first image will be assigned a score of 1.0, the second 0.8, the third 0.6, the fourth 0.4, and the fifth 0.2, with a consistent difference of 0.2 between each ranking position.

Next, visual feature extraction is carried out using a ViT model. Each image is divided into small patches, which are then enriched with positional encodings to preserve the spatial layout of the original image. These position-aware patches are passed through a transformer encoder, producing an embedding that captures both global and local visual information.

In parallel, textual encoding is performed for both the image descriptions and the main `sentence` using pretrained language models such as BERT or RoBERTa. These models convert the input text—whether in English or Portuguese—into contextual embeddings that capture semantic nuances, including idiomatic meanings.

To integrate the visual and textual modalities, a Cross-Attention module is applied. This module allows the text embeddings to attend to relevant regions of the image embedding, effectively aligning the linguistic and visual representations. This multimodal fusion is crucial for understanding idiomatic expressions, as these often involve metaphorical or symbolic visual cues that need to be associated with their corresponding textual interpretations. In

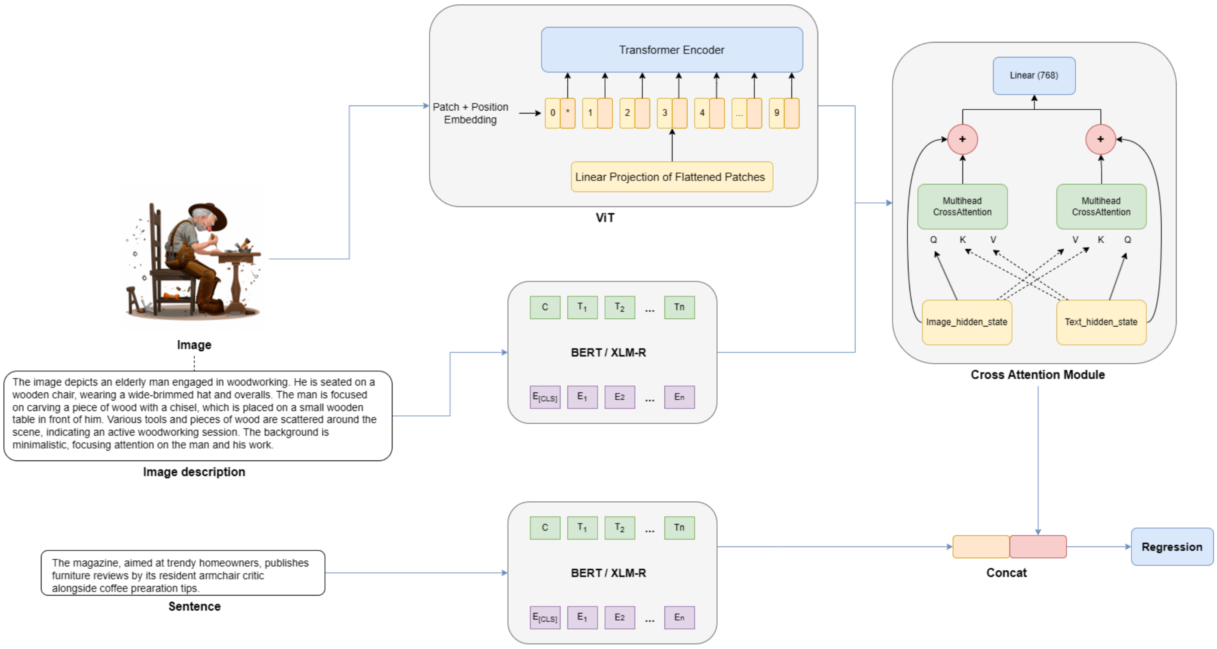

Figure 1: System architecture pipeline.

this case, we used a gated fusion mechanism based on multi-head attention with 8 heads and hidden dimensionality distributed across the heads. This mechanism dynamically balances the contribution of each modality, resulting in a unified multimodal representation that combines visual and textual context.

The resulting fused representation is then concatenated with the sentence embedding, producing a joint vector that encodes both visual and textual context. This combined representation is subsequently passed through a regression layer, which predicts a relevance score for each image in relation to the idiomatic meaning of the sentence. To evaluate the model's performance, we use Root Mean Square Error (RMSE) as the reference metric, as it effectively captures the deviation between the predicted and expected relevance scores.

The second approach is based on specific LLMs such as Llama-3.1-8B-Instruct (Dubey et al., 2024), taking advantage of their In-context learning capability and prompt engineering techniques such as zero-shot learning, to classify images according to their relevance in relation to idiomatic expressions. First, the model is prepared and a prompt (as shown in Listing 1) is generated with the sentence context and image descriptions. Then, the model evaluates each image and generates a similarity score, which indicates how relevant each image is to the meaning of the idiomatic expression. Finally, the images are sorted according to the score obtained and the results are saved in a file.

4 Experimental setup

For Subtask A, we have used only the training set provided by the organizers, which, after our preprocessing, is left with a total of 350 examples (of which, for each sentence, we have 5 related images).

For the first approach, different pre-trained language models have been tested, both multilingual and monolingual, such as BERT-base for English, BERTimbau-base (Souza et al., 2020) for Portuguese and, finally, XLM-RoBERTa for English and Portuguese. As for the ViT models, the "vit-base-patch16-224-in21k" (Wu et al., 2020) (Deng et al., 2009) model has been tested. The models are trained with 4 layers and 8 attention heads (num_heads) for the multimodal module, using a learning rate of `2e-5`, 20 epochs, and the RMSE metric as reference.

For the second approach, we use Llama-3.1-8B-Instruct. The parameters set for inference are: `do_sample=False`, which disables random sampling, ensuring reproducibility without the need to set the parameters `temperature`, `top_p` and `top_k`. In addition, a limit of `max_new_tokens=256` is set, ensuring that text generation does not exceed 256 tokens, optimizing model consistency and relevance in the image classification task.

5 Results

Table 1 presents the results obtained using three different approaches for the classification of images associated with idiomatic expressions: the BERT + ViT model, the XLM-RoBERTa + ViT model and the Llama-3.1-8B-Instruct model. The results were evaluated on two datasets: the Test set and the Extended eval set, both in English and Portuguese. The metrics used for the evaluation were Accuracy and Discounted Cumulative Gain (DCG) score.

In the test set, the approach of using Llama-3.1-8B as the classification model achieved the best results in terms of accuracy (0.4) and DCG score (2.677), standing out as the most effective approach. In comparison, the BERT + ViT model achieved an accuracy of 0.266 and a DCG score of 2.337, while the XLM + ViT model achieved an accuracy of 0.133 and a DCG score of 2.232. For the extended evaluation set, the results were similar. The LLM approach achieved an accuracy of 0.24 and a DCG score of 2.520, again standing out for its superior performance. The BERT + ViT model had an accuracy of 0.21 and a DCG score of 2.348, while the XLM + ViT model achieved an accuracy of 0.24 and a DCG score of 2.372.

In the Portuguese test set, the XLM + ViT model outperformed, achieving the highest accuracy (0.384) and the best DCG score (2.572). In comparison, the BERT + ViT model achieved an accuracy of 0.308 and a DCG score of 2.475, while the Llama model had an accuracy of 0.231 and a DCG score of 2.444. In the extended evaluation set in Portuguese, the BERT + ViT model was the best in terms of accuracy (0.236) and DCG score (2.387), while the LLM approach achieved an accuracy of 0.181 and a DCG score of 2.362. The XLM + ViT model, although obtaining a lower accuracy (0.182), achieved a DCG score of 2.331.

The results obtained show that the Llama model proved to be the most robust approach in English, achieving the best performance in terms of accuracy and DCG score in the test set. However, in Portuguese, the XLM + ViT model excelled in accuracy and DCG score in the test set, while the BERT + ViT model led the extended evaluation set.

Table 2 shows the results obtained using the Llama-3.1-8B approach developed by *UMUTeam*, as well as our position in the official Subtask A ranking.

On the English test set, the *PALI-NLP* team achieved first place with an accuracy of 0.933 and a DCG score of 3.581, followed by *dutir914* in second place with similar results. In this case, our approach based on Llama-3.1-8B achieved an accuracy of 0.4 and a DCG score of 2.677, placing us 17th in the test dataset ranking and 12th in the extended eval set. As for the Portuguese test set, the results were significantly different. The *HiTZ-Ixa* team led with a perfect precision of 1 and a DCG score of 3.505, followed by *dutir914* and *Zhoumou*. Using the XLM + ViT model, we achieved an accuracy of 0.384 and a DCG score of 2.572, which allowed us to reach 9th place in the test set and 8th place in the extended eval set.

It is important to note that the official evaluation system used by the organizers prioritized the model that showed the best performance in English, which resulted in a relatively low position for UMUTeam in the Portuguese ranking with the Llama-3.1-8B model. However, if we had employed the XLM + ViT model as the main approach for Portuguese, the results would have been considerably better, significantly improving our position in the Portuguese ranking.

Table 1: Results of different approaches: BERT + ViT (Approach 1), XLM + ViT (approach 2), and Llama-3.18b (Approach 3). Accuracy (ACC) and Discounted Cumulative Gain (DCG) scores are reported for the test (T) and the extended eval dataset (E)

#	ACC-T	DCG-T	ACC-E	DCG-E
English				
1	0.266	2.337	0.21	2.348
2	0.133	2.232	**0.24**	2.372
3	**0.4**	**2.677**	**0.24**	**2.520**
Portuguese				
1	0.308	2.475	**0.236**	**2.387**
2	**0.384**	**2.572**	0.182	2.331
3	0.231	2.444	0.181	2.362

6 Conclusion

Our participation in Subtask A of the AdMIRe shared task (SemEval 2025) focused on the complex challenge of interpreting idiomatic expressions using visual and contextual clues. We explored two distinct approaches: a multimodal model that fuses textual embeddings from BERT and XLM-RoBERTa with visual features from Vision Transformers, and an in-context learning ap-

Table 2: Official ranking of Subtask A for English and Portuguese. Accuracy (ACC) and Discounted Cumulative Gain (DCG) scores are reported for the test (T) and the extended eval dataset (E)

Team	Rank-T	Rank-E	ACC-T	DCG-T	ACC-E	DCG-E
			English			
PALI-NLP	1	1	0.933	3.581	-	-
dutir914	2	3	0.933	3.45	0.72	3.219
AlexUNLP	3	5	0.933	3.523	0.83	3.426
...	-	-	-	-	-	-
UMUTeam (Llama-3.1)	17	12	0.4	2.677	0.24	2.52
			Portuguese			
HiTZ-Ixa	1	7	1	3.505	0.454	2.821
dutir914	2	2	0.923	3.574	-	-
Zhoumou	3	3	0.923	3.425	0.690	3.061
...	-	-	-	-	-	-
UMUTeam (XLM + ViT)	9	8	0.384	2.572	0.182	2.331

proach utilizing the Llama-3.1-8B model for image classification.

Our experimental results reveal that while Llama-3.1-8B outperforms other models in English, the XLM + ViT model demonstrates superior accuracy and DCG scores in Portuguese. This highlights the importance of language-specific strategies for idiom interpretation. Furthermore, the success of the multimodal approach underscores the value of integrating visual and contextual information to better capture the figurative meanings of idiomatic expressions.

These findings contribute to the broader field of natural language understanding, particularly in enhancing machine comprehension of non-literal language. Future work will explore more advanced fusion techniques and investigate the role of cultural context in idiom interpretation to further improve model performance across languages.

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Fund for Regional Development (ERDF)-a way to make Europe, and the research project "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

References

Samuel A Bobrow and Susan M Bell. 1973. On catching on to idiomatic expressions. *Memory & cognition*, 1:343–346.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

María del Pilar Salas-Zárate, Giner Alor-Hernández, José Luis Sánchez-Cervantes, Mario Andrés Paredes-Valverde, Jorge Luis García-Alcaraz, and Rafael Valencia-García. 2020. Review of english literature on figurative language applied to social networks. *Knowledge and Information Systems*, 62(6):2105–2137.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Daniel García-Baena, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, and Rafael Valencia-García. 2023. Hope speech detection in spanish: The lgbt case. *Language Resources and Evaluation*, 57(4):1487–1514.

José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2023. Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english. *Mathematics*, 11(24).

Wei He, Thomas Pickard, Maggie Mi, Dylan Phelps, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Admire - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations*.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive Science*, 4(2):195–208.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES - Computer Modeling in Engineering and Sciences*, 140(3):2849–2868.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Alejandro Salmerón-Ríos, José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2024. Fine grain emotion analysis in spanish using linguistic features and transformers. *PeerJ Computer Science*, 10.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision.