Helsinki-NLP at SemEval-2022 Task 2 : A Feature-Based Approach to Multilingual Idiomaticity Detection

Itkonen, Sami

The Association for Computational Linguistics
2022-07-11

Itkonen , S , Tiedemann , J & Creutz , M 2022 , Helsinki-NLP at SemEval-2022 Task 2 : A Feature-Based Approach to Multilingual Idiomaticity Detection . in G Emerson , N Schluter , G Stanovsky , R Kumar , A Palmer , N Schneider , S Singh & S Ratan (eds) , Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) . The Association for Computational Linguistics , Stroudsburg , pp. 122-134 , International Workshop on Semantic Evaluation , Seattle , Washington , United States , 14/07/2022 .

http://hdl.handle.net/10138/346656

Helsinki-NLP at SemEval-2022 Task 2: A Feature-Based Approach to Multilingual Idiomaticity Detection

Sami Itkonen and **Jörg Tiedemann** and **Mathias Creutz**
Department of Digital Humanities
Faculty of Arts
University of Helsinki
Finland
{firstname.lastname}@helsinki.fi

Abstract

This paper describes the University of Helsinki submission to the SemEval 2022 task on multilingual idiomaticity detection. Our system utilizes several models made available by HuggingFace, along with the baseline BERT model for the task. We focus on feature engineering based on properties that characterize idiomatic expressions. The additional features lead to improvements over the baseline and the final submission achieves 15th place out of 20 submissions. The paper provides an error analysis of our model including visualisations of the contributions of individual features.

1 Introduction

We participated in the SemEval 2022 Task 2 (Tayyar Madabushi et al., 2022) Subtask A, zero-shot[1] setting: classification of a sentence containing a potentially idiomatic two-word multiword expression (MWE) as idiomatic or literal. The task provided four data sets[2] for English, Portuguese and Galician. Each MWE was represented by multiple example sentences, accompanied by the context (previous and next sentences). Each MWE could be always idiomatic, always literal or anything in between. Table 1 shows examples for both idiomatic (0) and literal (1) cases. Expanded examples (with context) are shown in Table 6 in the Appendix.

[1]The MWEs in the test data do not appear in the training data.

[2]Training, development, evaluation and test sets, with Galician only appearing in the final test set.

The motivation for our approach is testing linguistically motivated features that reflect important properties of idioms, such as non-compositionality, non-substitutability, non-literal-translatability and affectiveness (see chapter 2) and to see whether pre-trained models can be helpful for capturing these features. Our system uses a combination of models: BERT fine-tuning (Tayyar Madabushi et al., 2021), sentence embeddings (Reimers and Gurevych, 2019) and a feature model based on the above idiomatic properties.

2 Background and Related Work

The detection and analysis of idiomaticity has a rich history in the literature. An important property of idioms is non-compositionality (that is, the meaning of the expression does not correspond to the combination of the meaning of its components). (Peng et al., 2014; Constant et al., 2017; Gantar et al., 2018) Related to it are non-substitutability (components cannot be substituted with their synonyms) (Farahmand and Henderson, 2016; Senaldi et al., 2016; Constant et al., 2017) and non-literal-translatability (Constant et al., 2017).

Idioms tend to be semantic outliers (Feldman and Peng, 2013; Peng et al., 2014; Salton et al., 2016) in the sense that they violate the lexical cohesion of the surrounding discourse. They are also known to be more affective (either positive or negative) (Peng et al., 2014) than literal expressions.

In addition to being relatively fixed lexically (non-substitutability), idioms often exhibit lack of

Label	Target
0	He was not a *blue blood* jurist issuing judicial decisions that nobody understood affecting people and corporations that nobody knew.
1	The *blue blood* of the fossil-like creature is the only natural source of limulus amoebocyte lysate, a clotting agent that is used to test batches of injectable drugs for bacterial contamination that could cause fever, organ damage and even death.

Table 1: Idiomatic (0) and literal (1) examples from the zero_shot setting of training set for English MWE *blue blood*, which can be interpreted as either idiomatic or literal.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 122 - 134

syntactic and/or morphological variability (Peng et al., 2014; Constant et al., 2017). In general, quantifying any variability has traditionally required obtaining frequencies of the variants from a full corpus, as done by Inurrieta et al. (2020). However, as we only have a small number of examples for each idiom, these properties are not modeled in our approach.

Compositionality and substitutability are often tested with techniques like backtranslation and mask filling tasks. **Backtranslation** involves translating text to another (i.e. pivot) language and translating it back (Sennrich et al., 2016; Edunov et al., 2018), and it has often been used for paraphrasing and data augmentation. Backtranslations have also been used for idioms in related work, see, e.g., Moirón and Tiedemann (2006); Bahar Salehi and Baldwin (2018).

Mask filling (Zhu et al., 2019; Donahue et al., 2020) is closely related to the cloze task (Taylor, 1953), where the objective is to predict a word missing from an expression. Mask filling has lately been made easier as modern languages models such as BERT (Devlin et al., 2019) and its derivatives are themselves so-called Masked Language Models (MLM). Mask filling can be useful for testing substitutability in context (Karidi et al., 2021; Zhu and Bhat, 2021).

3 System Description

Our submission[3] considers three models: the baseline BERT model provided by the task authors (Tayyar Madabushi et al., 2021), sentence embeddings with sentence-transformers (Reimers and Gurevych, 2019) and a feature model based on idiomaticity features. All our components rely on existing models and tools that have been integrated into the transformers library provided by HuggingFace (Wolf et al., 2020).

The final classification combines information from two components (either fine-tuned BERT + feature model or sentence embeddings + feature model). The result will be taken from the model which has the higher label probability[4]. See Figure 1 for an overview.

We compare different variants of the system with the performance of individual features and various baselines.

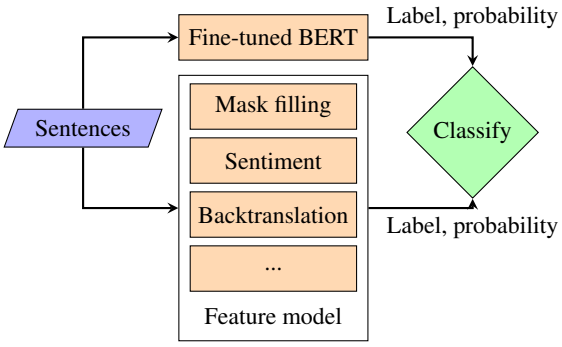


Figure 1: Basic classification procedure for the fine-tuned BERT + feature model combination. The feature model combines information from a number of HuggingFace models. Each model independently produces a label and associated probability. The label is by default taken from the model that has a higher probability. See Chapter 3.4 for the detailed classification procedure.

3.1 Fine-tuned BERT

The baseline model provided by the task organisers (Tayyar Madabushi et al., 2021) is based on BERT (Devlin et al., 2019). We build three variants: a) multilingual model (*bert-base-multilingual-cased*) for all languages (equivalent to the provided baseline), b) English model (*bert-base-cased*) for English data and multilingual model for non-English (trained with all data, including English) and c) same as case b, but multilingual model trained only with non-English data. The BERT model was fine-tuned with the training data, with the development set used for validation.

3.2 Sentence Embeddings (sbert)

Sentence embeddings can be used as an alternative baseline. We apply the *distiluse-base-multilingual-cased-v1*[5] model provided by HuggingFace and use the *sentence-transformers* python module[6]. The training procedure adopts the approach used by Tayyar Madabushi et al. (2021) by appending the MWE to the target sentence before training, as they found it to improve performance[7]. Logistic regression is used to train a classifier on top of the sentence embedding that we obtain from sbert.

[3] Implementation and details are available at https://github.com/dustedmtl/semeval2022.

[4] While both logistic regression and BERT models produce probabilities, the values aren't necessarily consummerate as BERT seems a lot more confident about the results.

[5] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[6] https://www.sbert.net

[7] This is not, however, equivalent to their methodology where the MWE is treated as a single token according to the "idiomatic principle" (i.e. stored as a single token in the mental lexicon) (Hashempour and Villavicencio, 2020).

Target	Label	Top terms	Top score	Hassub	Short	Found Idx	Found Score
There are several theories behind the origin of the term "*Double Dutch*."	0	., -, ..., s, *man*	0.008	False	3	10	-1.000
Além de ter sido um fracasso de bilheteria e crítica, o filme acabou marcado pelos seus *efeitos especiais*, principalmente ao antropomorfizar os gatos, que, bem, ficam um pouco bisonhos.	0	*erros*, personagens, **efeitos**, problemas, animais	0.540	True	0	3	0.092

Table 2: Training set substitution examples. In the first row, most of the suggestions are too short and no valid lexical substitute is found. The second example finds a component of *efeito especial* in plural form. The *Top terms* column shows the entry corresponding to *Top score* in *italics* and the one for *FoundIdx/Score* (if found) in **bold**. The above-zero scores represent the output from the mask-filling pipeline.

Note that we do not use the context sentences in this approach in any way.

3.3 Idiomaticity Features

Idiomaticity features are extracted using a number of HuggingFace pipelines and pre-trained models (see details below) for lexical substitution, sentiment analysis and backtranslation. Additionally, semantic outliers and surface-form-based boolean features are calculated. The training and classification with the feature model is done with logistic regression again, with boolean values converted to integers (True = 1 / literal, False = 0 / idiomatic).

3.3.1 Lexical Substitution

Because of the limited lexical variability and non-compositional nature of idiomatic expressions, it should be more difficult to find lexical substitutes for them, or their parts, than for literal expressions.

Our lexical substitution model utilizes the huggingface *fill-mask*[8] pipeline with the *xlm-roberta-base*[9] model. The pipeline will output a ranked list of top substitutions along with their scores[10]. Three different masks are used: one for masking the whole MWE (e.g. the expression *panda car* is replaced with <*mask*>)[11]), another for masking the first term (<*mask*> *car*) and a third one for masking the second term (*panda* <*mask*>)[12].

We obtain the top five candidates (individual words) from the pipeline. We are interested in two things: 1) how difficult it is to get a substitute in general, and 2) how difficult it is to get the *correct* substitute. The former reflects non-substitutability and the latter non-compositionality. A valid general substitute must only contain word characters and be at least three characters long. No other checks are made (such as whether the word class is correct or that the candidate is a synonym). A valid lexical substitute will additionally need to (case-insensitively) match either component of the MWE as it appears in the Target sentence. Inflected forms of the components are found by progressively stemming the component(s) with a regular expression- Additional tweaks are required for Portuguese because of orthographic variation (see Table 7 in the Appendix for examples).

The features that are generated are described below. Substitutions from masking individual terms are only used for the *Top score 1/2* and *FS/SS* features; all other features are derived from replacing the whole expression. Table 2 shows two examples for the features, with more examples in Table 8 in the Appendix.

Hassub Boolean feature: True when a valid lexical substitute is found, False otherwise.

Top score, Top score 1, Top score 2 The score of the top candidate, from replacing the whole expression, first term and second term, respectively. These features are a proxy for general (non-)substitutability.

Short, FS, SS The number of candidates that are too short (less than three characters) from masking the whole expression, first term and second term. The reasoning is that a lack of good suggestions is another proxy for non-substitutability.

FoundScore The score of the first valid lexical substitute [0-1], -1 otherwise. This one re-

[8] https://huggingface.co/tasks/fill-mask

[9] https://huggingface.co/xlm-roberta-base

[10] The mask-filling pipeline documentation doesn't explicitly state what the scores represent, but it's likely to be probability.

[11] The mask token is taken from the underlying model, which in this case is <*mask*>.

[12] The pipeline (by default) does not support using multiple mask tokens, so replacing the MWE with <*mask*> <*mask*> is not possible.

flects non-compositionality: replacement of the MWE with one of its components.

FoundIdx The index [1-5] of the first valid lexical substitute, 10 otherwise.

Note that the *FoundScore* and *FoundIdx* features essentially mix categorical and numeric values, which may reduce their usefulness. Additionally, *Top terms*, *Top terms 1* and *Top terms 2* are recorded from the mask filling process.

3.3.2 Sentiment Analysis

The affection feature is based on sentiment classification using the *cardiffnlp/twitter-xlm-roberta-base-sentiment*[13] model for predicting positive, negative or neutral sentiment. The neutral probability [0-1] is used as the value for the feature **Sentiment**.

3.3.3 Backtranslation

The target sentence is translated to another language (Portuguese for English, English for Portuguese and Galician) and then back-translated to the original language with the OPUS-MT (Tiedemann and Thottingal, 2020) models *opus-mt-en-roa* and *opus-mt-roa-en*[14]. The rationale is that idiomatic expressions exhibit non-compositionality and as such are less likely to be backtranslated correctly. The logic for locating the expression is the same that was used for lexical substitution: the exact form is required, with allowances for variations in Portuguese orthography. The value for the feature **Trans** is True if it is found and False otherwise. Table 7 in the Appendix shows a number of backtranslated examples.

3.3.4 Semantic Outliers

To measure semantic coherence, sentence embeddings are retrieved from sentence-transformers for the sentences/expressions. The value of the feature is the cosine similarity between the two.

Prevdiff Cosine similarity between the Previous and Target sentences.

Nextdiff Cosine similarity between the Next and Target sentences.

MWEdiff Cosine similarity between the MWE and the Target sentence.

3.3.5 Surface-form features

Based on data exploration, two additional surface features are used:

Quotes True if the MWE is enclosed in quotes, in which case it is more likely to be idiomatic.

Caps True if the MWE is capitalized (Camel Case). This is more likely to be a Proper Noun.

Table 9 in the Appendix shows examples for these features.

3.4 Final Classification

With the exception of simple baselines and majority voting classifier, the final classification is done by combining two components. For each prediction, the result will be taken from the model which has the higher probability. A number of ablation tests were run for the feature model with the development set to select the best set of features. In the end, all features except *Top score 1*, *FS* and *MWEdiff* were retained (where *features* means the bolded items in Chapters 3.3.1 through 3.3.5).

We also added a final post-correction step based on the results observed during development: the boolean features may (potentially) override the label. There are two modes for this: the first one will force the label unconditionally, the second one will force it only if the models disagree (*agree*).

The potential idiomatic features are *Quotes* and *!Trans* (*Trans* == False, that is, a mistranslation). Potential literal features are *Hassub* and *Caps*. Literal features take precedence, so if an expression is both quoted and capitalized, it is considered literal.

4 Results

4.1 Experimental Setup

Four data sets were released by the task administrators: training and developments sets, for which gold labels were provided; an evaluation set without gold labels (for which classification results could be obtained from the competition website) and a blind test set. The training set had more idiomatic (56%) and the development set more literal (54%) sentences.

The label is overwhelmingly likely to be 1 (literal) when the surface feature *Caps* == True (see

[13] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[14] https://huggingface.co/Helsinki-NLP/opus-mt-en-roa and https://huggingface.co/Helsinki-NLP/opus-mt-roa-en. The models were chosen out of convenience, as only two models are required for translating between the languages to either direction.

Configuration	F1	EN	PT
Hassub	0.551	0.535	0.547
Trans	*0.597*	0.549	*0.615*
Sentiment	0.542	0.571	0.486
Majority class	0.545	*0.609*	0.564
Sentence transformers	0.614	0.635	0.536
+ features	*0.713*	*0.735*	*0.629*
BERT baseline	0.694	0.705	0.612
+ ml1: PT from full model	0.721	0.760	0.612
+ ml2: PT from separate model	0.725	0.760	0.590
+ features	0.715	0.716	0.656
+ ml1 + features	0.733	0.751	0.666
+ ml1 + features, agree	0.731	0.754	0.656
+ ml1 + features + trans	**0.751**	0.762	**0.694**
+ ml1 + features + trans, agree	0.750	**0.767**	0.683
+ ml2 + features + trans, agree	0.742	**0.767**	0.642
Majority voting + trans	*0.724*	0.743	*0.645*
Majority voting + trans, agree	0.723	*0.746*	0.633

Table 3: Results for the development set. Sections in order: baselines; combinations with sentence embeddings; BERT fine tuning models; majority voting classifiers. For BERT models, ml1 uses English model for English and and multilingual model for Portuguese (trained on all data), while ml2 is only trained with Portuguese data. Best/second best results are **bolded**/underlined, while the best result for each section is in *italics*.

Configuration	F1	EN	PT
Sentence transformers	0.558	0.579	0.500
+ features	0.646	0.655	**0.615**
BERT baseline	0.702	0.760	0.566
+ ml1	0.723	0.791	0.578
+ features	0.714	0.779	0.577
+ features + trans	0.671	0.695	0.591
+ features, agree	0.723	0.791	0.578
+ ml1 + features	0.720	0.794	0.577
+ ml1 + features, agree	**0.725**	**0.800**	0.578

Table 4: Results for the evaluation set. The feature model provides less improvement over the baseline. For Portuguese, the sbert+feature model combination outperforms all BERT-based variants. Best/second best results are **bolded**/underlined.

Language	F1
English	0.752
Portuguese	0.694
Galician	0.499
Total	0.663

Table 5: Official results for the test set.

Figure 3 in the Appendix). We also found the features *Hassub* and *Quotes* to be useful, so they are used in all cases involving the feature model.

4.2 Development and Evaluation Sets

Results for the development set are shown in Table 3 for various baselines, sentence-transformers-based models and BERT-based models. Baselines for the boolean *Hassub* and *Trans* are taken directly from the feature: True=1, False=0, while for *Sentiment* above-mean scores are considered literal. *Majority class* assigns the majority label (literal) for all sentences.

The sentence embeddings + feature model yields better results than the base BERT model, but in general fine-tuning BERT is much better than using sentence embeddings as a fixed feature. For the BERT-based models[15], using an English-only model for English improves results, as does using the *!Trans* boolean feature. Using the boolean features only when the models disagree (*agree*) does not seem to have much impact. As Figure 4 in the Appendix shows, the BERT-based models are more likely to label a sentence as literal. Finally, the majority voting classifier (using the majority label from all three classifiers) fares worse than BERT+feature models.

[15]The baseline, multilingual 1 and 2 (ml1 and ml2) configurations refer to variants a-c in section 3.1.

The results for the evaluation set (in Table 4 are largely similar to those for the development set, except for two things: 1) the *!Trans* feature is detrimental to English and somewhat helpful for Portuguese and 2) boolean features should be used only when the underlying models disagree. In the end, using the feature model with BERT only slightly improves the result (0.725 > 0.723). Additionally, sentence embeddings + feature model approach outperforms BERT-based models for Portuguese.

4.3 Test Set

The test predictions were generated with the setup that produced the best overall results for the evaluation set: different BERT models for English and non-English combined with the feature model with boolean features *Hassub*, *Quotes* and *Caps* (only when the models disagree about the label). The official test results in Table 5 show that the results for Galician are not great - roughly on the level of random chance[16]. The official baseline isn't much better[17], likely due Galician being a low-resource language and lacking training data for the pre-trained models that were used. For English

[16]Without knowing the true labels, we assume a 50/50 split.

[17]https://sites.google.com/view/semeval2022task2-idiomaticity/baselines

and Portuguese, the results are similar to the best results for the development set.

Regarding specific features, the results lend support to the idea that idioms are more affective, thus sentiment analysis can be useful for detecting idiomaticity (see Figure 2 and Figure 6 in the Appendix). The exception here seems to be Galician, which is probably because the sentiment model is based on tweets. However, it is easier to get a lexical replacement for Galician (see Figure 5a in the Appendix). It may be possible that the Galician test sentences use simpler language - relatively speaking.

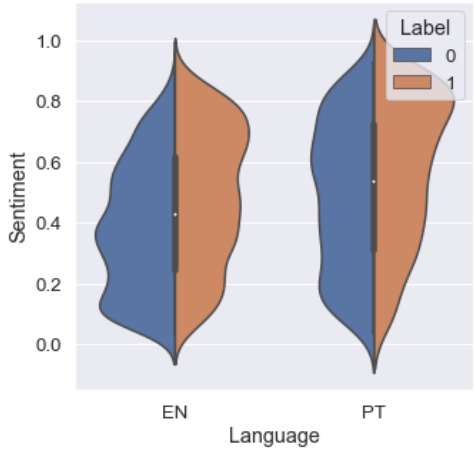


Figure 2: Violin plot for sentiment per language for the training set. The skewed sentiment distribution shows that the label is more likely to be literal (on average) for both English and Portuguese, if the sentiment score is higher (neutral sentiment). However, this feature alone is not sufficient for good performance.

Using the boolean features on top of the classifier models can be a bit of a hit-and-miss: what works with one dataset may be detrimental with another. Specifically, the *!Trans* feature worked well on the development set, but not on the evaluation set, and the *Hassub* feature worked on both of these sets, but not on the test set. In other words, the boolean features may make the model less robust.

Ablation studies performed after the official end of the competition confirm that using the *Hassub* feature for the test set was not a good strategy. Furthermore, a feature-only model (without sentence embeddings or BERT) outperformed the combined model, with the best results achieved by using the combined model for English and feature-only model for Portuguese and Galician. Nevertheless, even these results do not come close to the best models of the competition.

For detecting semantic outliers, the approach used in this paper (similarity based on sentence-transformers embeddings) appears to be too simple. More refined methods, such as those measuring lexical cohesion (Sporleder and Li, 2009) would be required.

5 Conclusions

Our system combines a feature model based on a number of idiomaticity features with a BERT transformer classifier. The feature model achieves competitive results compared to the reportedly strong baseline (Tayyar Madabushi et al., 2022), although it does not fare nearly as well as the best systems that competed in the subtask. Unsurprisingly, most of the features work best for English, whether or not the underlying BERT model is multilingual or not.

The work shows that a classification system utilizing idiomatic properties such as non-compositionality, non-substitutability and affectiveness can be implemented with readily available transformer APIs.

Another idea for future work is to improve the back-translation test by combining a "good" forward translation model (i.e. one that tends to properly treat idiomatic expressions) with a "bad" back-translation model (i.e. one that tends to produce literal translations). The latter could also be done by forcing component-wise translations in the back-translation step to reveal non-compositionality of the expression.

Acknowledgements

The work in this paper was supported by the project Behind the Words funded by the Academy of Finland. Computational resources were provided by CSC, the Finnish IT Center for Science. We also thank the anonymous reviewers for their substantive comments.

References

Paul Cook Bahar Salehi and Timothy Baldwin. 2018. Exploiting multilingual lexical resources to predict mwe compositionality. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, page 343–373, Berlin, Heidelberg.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword

expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

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.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Meghdad Farahmand and James Henderson. 2016. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 61–66, Berlin, Germany. Association for Computational Linguistics.

Anna Feldman and Jing Peng. 2013. Automatic detection of idiomatic clauses. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'13, page 435–446, Berlin, Heidelberg. Springer-Verlag.

Polona Gantar, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso. 2018. Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*, 32(2):138–162.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, Online. Association for Computational Linguistics.

Uxoa Inurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2020. Learning about phraseology from corpora: A linguistically motivated approach for Multiword Expression identification. *PLoS ONE*, 15(8 August):1–18.

Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend, and Vivek Srikumar. 2021. Putting words in BERT's mouth: Navigating contextualized vector spaces with pseudowords. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10300–10313, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Begoña Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Giancarlo D. Salton, Robert J. Ross, and John D. Kelleher. 2016. Idiom token classification using sentential distributed semantics. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 1:194–204.

Marco Silvio Giuseppe Senaldi, Gianluca E. Lebani, and Alessandro Lenci. 2016. Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 21–31, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

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)*. 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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text Infilling.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic Phrase Detection by Masked Language Model. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168.

A Appendix

A number of tables and figures are presented here.

Table 6 shows samples from the training data. Tables 7, 8 and 9 list feature examples for substitution, backtranslation and Quotes/Caps surface features.

Figure 3 plots the boolean features *Hassub*, *Quotes*, *Caps* and *Trans* against the literal and idiomatic labels for the training set. Figure 4 demonstrates the differences in the labeling done by fine-tuned BERT and the feature model.

Figure 5 shows counts for boolean features for each language and data set. Illustrations for sentiment scores for all languages and datasets are shown in Figure 6.

Label	Previous	Target	Next
0	Heading outside (even just for a couple of minutes) or doing mundane things like brushing your teeth and making the bed can help your mind accept the fact that yes, alas, you are awake now.	Whether you're a *night owl* or early bird, though, try to make sure you're not diving right onto your phone.	Your morning will start calmer if you don't dive right into work emails and scrolling.
1	LCG asks that Monday customers put garbage and recycling carts at the curb for collection Tuesday morning.	In addition, the Lafayette Transit System office will be closed Monday, and there will be no Daytime, *Night Owl* or Paratransit bus service Monday.	Bus and paratransit services will resume regular schedules Tuesday.
0	I practiced before him in court and stood beside him on Canal Street during Endymion.	He was not a *blue blood* jurist issuing judicial decisions that nobody understood affecting people and corporations that nobody knew.	His blood was red with a little Irish green thrown in.
1	The American horseshoe crab has outlived the dinosaurs and survived four mass extinction events, but its population has been devastated in recent years, partly due to harvesting for biomedical production.	The *blue blood* of the fossil-like creature is the only natural source of limulus amoebocyte lysate, a clotting agent that is used to test batches of injectable drugs for bacterial contamination that could cause fever, organ damage and even death.	The crabs are fished from the oceans, taken to a lab to have about 30% of their blood harvested, then released back into the wild.

Table 6: Idiomatic (0) and literal (1) examples from the training set for English MWEs *night owl* and *blue blood* in the zero_shot setting. For *night owl*, the second example is considered literal as the MWE refers to a company name (Proper Noun).

MWE	Target	Label	BT	Trans
double dutch	Since settlers from other areas of the world could not understand the songs, they labeled the activity “*Double Dutch*.”	1	As settlers from other parts of the world could not understand the songs, they labeled the activity "*Double Dutch*".	True
double dutch	At 6,400gns, Auldhouseburn sold another by the same sire, and again in lamb to *Double Dutch*, to Northern Irish buyer, J. Cubbitt of Ballymena.	1	At 6,400gns, Auldhouseburn sold another by the same sire, and again in lamb to Duplo Dutch, to the Northern Irish buyer, J. Cubbitt of Ballymena.	False
círculo virtuoso	Com a segurança da imunização em massa e os números traduzindo sua eficácia, fica mais fácil para o americano médio sentir-se confiante em marcar sua próxima viagem, gerando um *circulo virtuoso* para o setor nos próximos meses.	0	Com a segurança da imunização em massa e os números traduzindo sua eficácia, torna-se mais fácil para o americano médio sentir-se confiante em marcar sua próxima viagem, gerando um *círculo virtuoso* para o setor nos próximos meses.	True
círculo virtuoso	Apesar de dizer que o Brasil está no caminho de um "*círculo virtuoso* na economia", o executivo do banco enxerga riscos internos e externos no horizonte da renda variável e, por isso, evita projeções de curto prazo.	0	Apesar de dizer que o Brasil está no caminho de um "*círculo virtuoso* na economia", o executivo do banco vê riscos internos e externos no horizonte da renda variável e, portanto, evita projeções a curto prazo.	True
amor-próprio	No novo livro, sobre *amor-próprio* e também validação social, Paula Cordeiro relata como sobreviver à era digital.	1	No novo livro, sobre o *amor próprio* e também a validação social, Paula Cordeiro relata como sobreviver à era digital.	True

Table 7: Backtranslation examples for the training set; *Target* is the original sentence, *BT* is the backtranslated one. The matching process occasionally requires some tweaks for Portuguese. In the third row, the Target contains the expression *círculo virtuoso* without an accent, while the last row shows the translation of *amor-próprio* separated with a space instead of a dash.

Target	Label	Top terms	Top score	Hassub	Short	Found Idx	Found Score
There are several theories behind the origin of the term "*Double Dutch*."	0	., -, ..., s, *man*	0.008	False	3	10	-1.000
Double Dutch also derives from the same era, Dutch seeming a strange and convoluted language hence Double Dutch meaning indescernible, mad and generally all round not on foreign speak.	0	It, *English*, This, **Dutch**, German	0.385	True	1	4	0.062
No vídeo divulgado nas redes sociais, é possível perceber que um som faz o casal olhar para o prédio da frente e ver o *efeito especial* da fumaça.	0	*som*, **efeito**, tamanho, aumento, ar	0.214	True	1	2	0.120
Os *efeitos especiais* são necessários em cenas de batalha, porém, a DC costuma abusar da técnica.	0	***efeitos***, equipamentos, personagens, dados, filmes	0.158	True	0	1	0.158

Table 8: Abbreviated substitution examples for the training set. The first two examples are for the English MWE *double dutch*, the last two for the Portuguese MWE *efeito especial*. In the first row, a substitution is not found and most of the suggested substitutions are too short, leading to a *Short* value of 3. In the second row, the fourth suggestion matches the MWE. For Portuguese, the expression *efeito especial* is found in singular form in the first example and in plural form in the second; the substitute suggestions must match the expression. The *Top terms* column shows the entry corresponding to *Top score* in *italics* and the one for *FoundIdx* and *FoundScore* (if found) in **bold**. The scores represent the output from the mask-filling pipeline.

MWE	Target	Label	Quotes	Caps
double dutch	*Double Dutch* also derives from the same era, Dutch seeming a strange and convoluted language hence Double Dutch meaning indescernible, mad and generally all round not on foreign speak.	0	False	True
double dutch	Since 1977 we have had a plethora of Foreign Ministers, to whom the subject of foreign affairs was *double Dutch*.	0	False	False
double dutch	At 6,400gns, Auldhouseburn sold another by the same sire, and again in lamb to *Double Dutch*, to Northern Irish buyer, J. Cubbitt of Ballymena.	1	False	True
night owl	The researchers said experience shows *"night owl"* patients with depression are less likely to recover and are more likely to commit suicide.	0	True	False

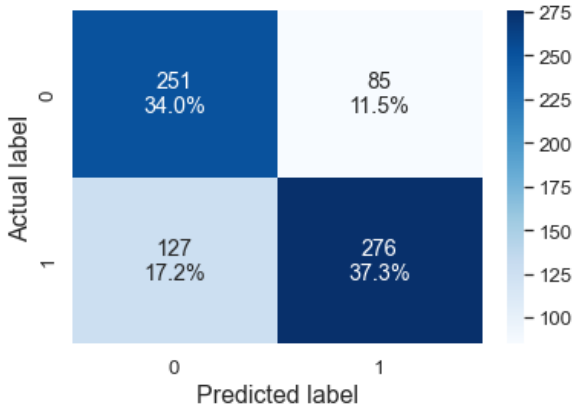
Table 9: Quotes/Caps examples for the training set. In the second row, *Caps* == False as both components of *double dutch* are not capitalized.

(a) All

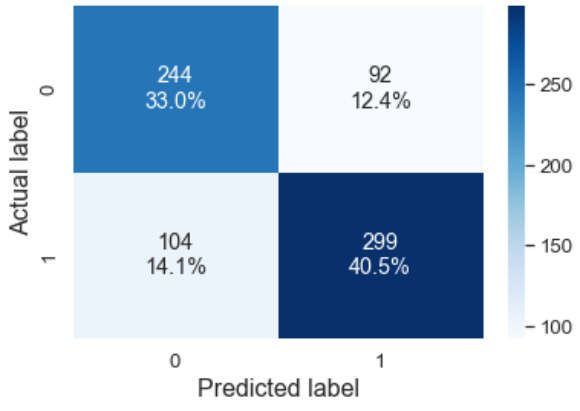
(b) English

(c) Portuguese

Figure 3: Boolean features vs label for the training set. The label is overwhelmingly likely to be literal if the MWE is Capitalized (*Caps* == True), while idiomatic label is more likely if the MWE is mistranslated (*Trans* == False). It is generally difficult to get a valid lexical substitute (*Hassub* == True).

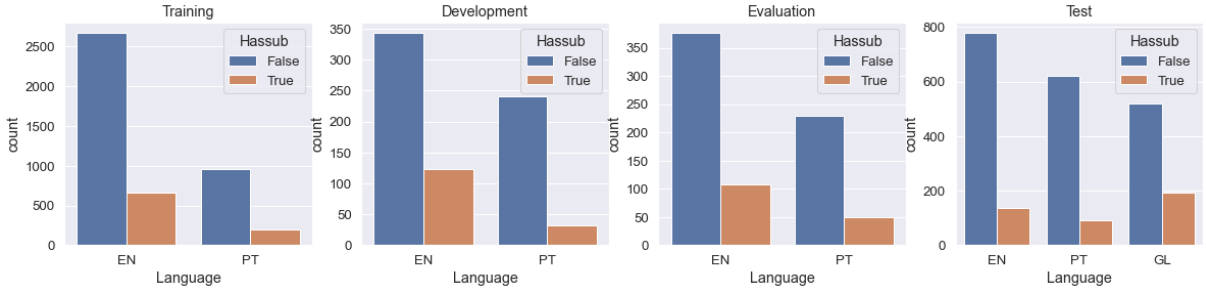


(a) Confusion matrix for the sbert+feature model.

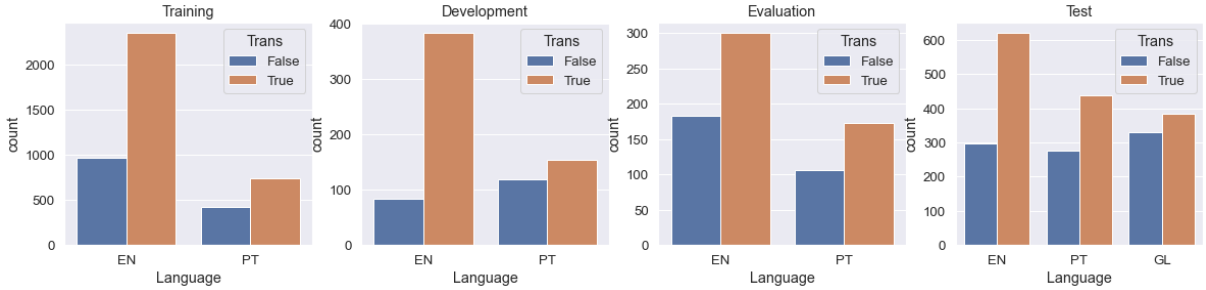


(b) Confusion matrix for the BERT+feature model.

Figure 4: Confusion matrices for the development set. The fine-tuned BERT model is more likely to classify the sentence as literal than the feature model.

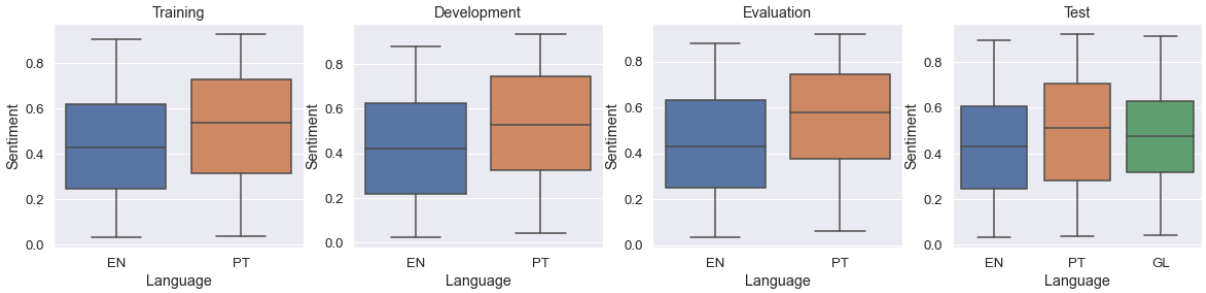


(a) Hassub



(b) Trans

Figure 5: Counts per feature, set and language. It is relatively easier to get a valid lexical substitute for Galician. Getting a correct backtranslation is harder for Portuguese than English, and harder still for Galician.



(a) Box plot

(b) Violin plot

Figure 6: Sentiment scores per set and language. The distributions are skewed for English and Portuguese, while the sentiment scores seem uninformative for Galician. Portuguese scores are generally higher - it is more difficult for the sentiment classifier to classify sentences as affective (either positive or negative).