Text Mining at SemEval-2024 Task 1: Evaluating Semantic Textual Relatedness in Low-resource Languages using Various Embedding Methods and Machine Learning Regression Models

Ron Keinan
Department of Computer Science, Jerusalem College of Technology, Lev Academic Center
21 Havaad Haleumi St., P.O.B. 16031, 9116001 Jerusalem, Israel
ronke21@gmail.com

Abstract

In this paper, I describe my submission to the SemEval-2024 contest. I tackled subtask 1 - "Semantic Textual Relatedness for African and Asian Languages". To find the semantic relatedness of sentence pairs, I tackled this task by creating models for nine different languages. I then vectorized the text data using a variety of embedding techniques including doc2vec, tf-idf, Sentence-Transformers, Bert, Roberta, and more, and used 11 traditional machine learning techniques of the regression type for analysis and evaluation.

1 Introduction

Semantic Textual Relatedness (STR), which involves determining the degree of semantic similarity or relatedness between two pieces of text, has emerged as a significant task within Natural Language Processing (NLP). This task holds significant relevance and importance across various applications, including information retrieval, question answering, and summarization. By accurately measuring the semantic relatedness between sentences, we can enhance the performance of many NLP systems and improve their overall effectiveness.

In this paper, we describe our participation in subtask 1-A of SemEval 2024, for STR of texts written in 9 languages: Algerian Arabic, Amharic, Hausa, Kinyarwanda, Moroccan Arabic, Marathi, Telugu, Spanish, and English. Our approach to solving the task was based on a previous study that dealt with a similar sentiment classification task (Keinan & HaCohen-Kerner, 2023), and was based on a comparison of different embedding methods and then a comparison between different regression classifiers. We compared the results of each classifier with other vectors and chose the model that provided the best results on the training dataset, in favor of classifying the proximity between the sentences in the test dataset.

The full description of task 1 in general and the subtasks, in particular, is given in Ousidhoum et al. (2024B), and the dataset is described in Ousidhoum et al. (2024A).

2 Background

2.1 Semantic Textual Relatedness

Semantic Textual Relatedness (STR) is pivotal in automatically assessing the semantic similarity or relatedness between pieces of natural language text, thereby offering insights into the underlying relationships between subjects(Hadj et al., 2020). STR facilitates the exploration of individuals' opinions on specific topics and enables actionable insights for future planning(Abdalla et al., 2023).

In an era marked by the proliferation of textual data across various platforms, STR serves vital purposes such as information retrieval, question-answering, and summarization. Despite the inherent complexities in STR, including nuances in language and the varying degrees of relatedness between texts, researchers are actively engaged in refining and advancing STR systems to achieve greater precision in measuring semantic textual relatedness.

Challenges abound both for computational algorithms and human evaluators in STR. Achieving accurate results in STR demands not only an understanding of linguistic context but also cultural context and specific domain knowledge(Gabrilovich & Markovitch, 2007).
Budanitsky and Hirst (2006) argued that relatedness is more general than similarity, as the former subsumes many different kinds of specific relations, including opposition, functional association, and others. They claimed that computational linguistics applications often

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 420–431
June 20-21, 2024 ©2024 Association for Computational Linguistics

require measures of relatedness rather than the more narrowly defined measures of similarity.

2.2 Semantic Textual Relatedness in Low Resources African Languages

Detecting STR in low-resource African and Asian languages poses an even greater challenge for several factors. In the realm of STR, tackling the scarcity of annotated data emerges as a significant hurdle, particularly concerning low-resource languages. Annotated data, crucial for training ML algorithms in STR, denotes text labeled with sentiments, like positive, negative, or neutral. This dearth of annotated data hampers the development of high-quality STR systems. ML algorithms thrive on ample data to discern patterns and make precise predictions. Consequently, STR systems tailored for low-resource African/Asian languages, lacking sufficient annotated data, often exhibit diminished performance and accuracy.

Moreover, the variability of sentiment expressions in low-resource African/Asian languages poses another formidable challenge. Unlike English, many languages boast a diverse palette of emotional expressions, complicating sentiment determination. Cultural nuances further compound this complexity, influencing the sentiment encoded within the text.

Furthermore, the scarcity of NLP tools and resources makes the task even harder. Text preprocessing, a crucial step in preparing data for SA, becomes arduous due to the limited availability of essential tools like stemming and lemmatization tailored for low-resource languages. This scarcity impedes effective text processing and hinders progress in developing robust STR systems for these languages.

Muhammad et al. (2022) embarked on an extensive research endeavor aimed at constructing a comprehensive database encompassing four resource-poor African languages. Their work stands out for its innovative contributions, including the development of stopwords databases and sentiment dictionaries tailored specifically for Nigerian languages.

Kelechi et al. (2021) ventured into training a multilingual language model exclusively on low-resource African languages. Their creation, AfriBERTa, spans eleven African languages, pioneering language models for four of these languages.

Dossou et al. (2022) introduced AfroLM, a multilingual language model trained from scratch on a staggering twenty-three African languages, employing a self-active learning framework. Their research highlights AfroLM's remarkable performance surpassing several multilingual pre-trained language models, including AfriBERTa, XLM-Roberta-base, and mBERT, across various downstream natural language processing tasks such as Named Entity Recognition (NER), Text Classification (TC), and Sentiment Analysis.

2.3 Text Preprocessing

Text preprocessing is crucial in NLP fields such as STR. In both general and social text documents, noise such as typos, emojis, slang, HTML tags, spelling mistakes, and repetitive letters often appear. Improperly preprocessed text can result in incorrect analysis outcomes.

HaCohen-Kerner et al. (2019, 2020) investigated the impact of all possible combinations of six preprocessing methods on TC in three datasets. The main conclusion recommended is always to perform a systematic variety of preprocessing methods, combined with many ML methods to improve the accuracy of TC.

2.4 Text Embeddings

Text embeddings are representations of textual data in a continuous vector space, enabling algorithms to process and analyze text effectively. These embeddings capture semantic and syntactic similarities between words or documents, facilitating various NLP tasks such as sentiment analysis, document classification, and information retrieval. We used 5 basic embedding methods: TF-IDF, Doc2Vec, mUSE, LSA, LDA, and 2 improved embedding methods – BERT and Sentence Transformers with a variety of models.

TF-IDF (Term Frequency-Inverse Document Frequency) represents the importance of a word in a document relative to a collection of documents. It calculates a weight for each word based on its frequency in the document and inverse frequency across all documents. Words with high TF-IDF scores are considered more informative for distinguishing documents(Ramos, 2003).

Doc2Vec, an extension of Word2Vec, generates fixed-length vectors for entire documents. It captures semantic information by training a neural network to predict the context of words within a document. Doc2Vec assigns a unique vector to each document, enabling comparison and clustering of documents based on their content(Lau & Baldwin, 2016).

mUSE (Multilingual Universal Sentence Encoder) is a pre-trained sentence encoder developed by Google Research. It maps variable-length text inputs into fixed-length vectors, capturing semantic similarity across multiple languages.

LSA (Latent Semantic Analysis) is a dimensionality reduction technique applied to large textual corpora. It analyzes the relationships between words and documents based on the co-occurrence of terms.

LDA (Latent Dirichlet Allocation) is a probabilistic generative model used for topic modeling. It assumes that documents are composed of multiple topics, each characterized by a distribution of words. LDA infers the underlying topic structure of a document collection and assigns a probability distribution over topics for each document.

BERT (Bidirectional Encoder Representations from Transformers), introduced by Google, employs a transformer architecture to capture bidirectional contextual information from text(Devlin et al., 2018). It consists of multiple layers of transformers, enabling it to understand the context of a word within a sentence based on both preceding and succeeding words(Chi et al., 2019).

Sentence-Transformers(ST), inspired by the success of BERT, extend its capabilities to encode entire sentences or paragraphs into fixed-length embeddings. Unlike BERT, which focuses on token-level representations, ST generates embeddings at the sentence level. These embeddings capture the contextual relationships between words within a sentence.

2.5 Task and Datasets Description

The SemRel Task 1-A is based on a collection of datasets in 9 different languages(Ousidhoum et al., 2024B). Each instance in the training, development, and test sets is a sentence pair. The instance is labeled with a score representing the degree of semantic textual relatedness between the two sentences. The scores can range from 0 (maximally unrelated) to 1 (maximally related. The size of the datasets is detailed in Appendix A. The official evaluation metric for this task is the Spearman rank correlation coefficient, which captures how well the system-predicted rankings of test instances align with human judgments.

3 System Overview

In our study, we implemented a systematic approach to enhance the learning process of our classifier. To augment the available training data, we merged the datasets of the training and development sets. This consolidation aimed to enrich the information on which our classifier is trained. Subsequently, we conducted experiments where each model was evaluated on both raw sentences and preprocessed sentences. The preprocessing steps included removing punctuation marks, numeric characters, and URL addresses, and converting text to lowercase.

At each stage of the learning process, we employed various text embedding methods to convert sentence pairs into vector pairs. These text embedding methods were pivotal in capturing the semantic relationships between sentences. Following the generation of vector pairs, we trained a regression model to learn the Semantic Textual Relatedness (STR) label between the vector pairs. The trained model was then tasked with predicting the STR level for unlabeled vector pairs present in the test set. Subsequently, we performed a comparative analysis of all results and selected the best-performing models for each language under investigation.

Furthermore, we evaluated the performance of eleven machine learning regressors to determine their efficacy in predicting the STR label. These regressors include:

Linear Regression: A basic regression model that models the relationship between independent and dependent variables linearly.

Ridge Regression: A regression model that uses L2 regularization to prevent overfitting.

Gradient Boosting Regressor: An ensemble learning technique that builds decision trees sequentially, each correcting the errors of the previous one.

AdaBoost Regressor: Another ensemble learning method that combines multiple weak learners to create a strong learner.

Support Vector Regressor (SVR): A regression algorithm that finds the hyperplane that best fits the data points while minimizing the error. SVM is a supervised learning algorithm that is used for classification and regression analysis(Cortes and Vapnik, 1995; Chang & Lin, 2011).

Stochastic Gradient Descent (SGD) Regressor: A linear model trained using stochastic gradient descent.

Bayesian Ridge Regression: A regression model that is based on the Bayes theorem (Kim et al., 2006), and assumes that features are conditionally independent given the target class, estimates the probabilities of each class and the probabilities of each feature given the class, and use it to make predictions.

Decision Tree Regressor: A regression model that partitions the data into subsets based on feature values.

Random Forest Regressor: An ensemble learning method that builds multiple decision trees and outputs the average prediction. (Breiman, 2001). It combines Breiman's “bagging” (Bootstrap aggregating) idea in Breiman (1996) and a random selection of features introduced by Ho (1995) to construct a forest of decision trees.

K Neighbors Regressor: A non-parametric regression model that predicts the output based on the average of the 'k' nearest neighbors.

MLP Regressor (Multi-layer Perceptron): A neural network model with multiple layers that learns complex data patterns. Inputs are received by the input layer, processed through the hidden layers, and produce the final output (Hassan et al., 2016).

Each regressor was evaluated based on its performance in predicting the STR label, providing insights into the effectiveness of different regression techniques in our task.

4 Experimental Setup

Our way of working was based on the train and dev datasets only. The goal was to train different models on the train dataset and select the best models according to the Spearman rank score (according to the competition requirement) on the dev dataset.

For all embedding methods(see Appendix B for details), we applied the following process. In the first step, for each language, converted the sentence pairs to vectors, using different embedding methods. Every method was checked twice – one with the original pair and one with a preprocessed pair. In total, for each language, we tested 5 classic embedding methods, 4 methods based on Sentence-Transformers, and 8 methods based on BERT.

That is, for each language different embedding methods were tested, once on raw text and once on pre-processed text, and for each of these methods we trained 11 regression models. We also trained additional BERT models for English, Spanish, Moroccan Arabic, and Algerian Arabic, so that in total we compared 3572 models (for all languages together), and at least 374 models for each language.

The following tools and information sources were utilized to apply these ML methods:

Python 3.8 programming language (Van Rossum & Drake, 2009),

Sklearn – a Python library for ML methods (Buitinck et al., 2013),

Numpy – a Python library for fast algebraic calculation (Harris et al., 2020),

Pandas – a Python library for efficient data analysis (McKinney, 2010),

TensorFlow – an open-source Python library for constructing ML-DL models (Abadi et al., 2015), and

Transformers – a Python library for natural language processing, offering pre-trained models based on transformer architecture (Wolf et al., 2020).

Hugging Face - provides a platform for data scientists to access and utilize cutting-edge models (Huggingface API, 2024).

5 Experimental Results

Table 1 presents the Spearman rank score of our models for task 1A. The table shows for each language the ideal model we received, its embedding method, whether it performed pre-processing, which regressor it used, what was the score we received in the training phase (distribution of train+dev in the ratio 20:80), what was the actual score we received after submission to the competition, and what was our position in the competition as well as what is the best result achieved. The full results can be seen in Appendix C.

It seems that vector assignment in BERT-based embedding methods was better than classical methods or Sentence Transformers library-based methods. This is probably due to the work that these models were massively trained on a lot of information, with the help of huge resources, and are therefore able to characterize vectors that optimally deliver the texts. Also, BERT models know how to characterize words with their context, and this may be a significant fact concerning the STR task.

In most languages, except Spanish and Kiryanwanda, a BERT model that is multilingual was better than a BERT model that was trained only on this or a similar language. This is a surprising figure as we were sure that a specific model would excel more reliably in this language. However, it seems that the models in low-resource languages are weaker and trained on less information compared to huge models from the multilingual genre.

Among the classical embedding methods, tf-idf seems to be the most successful method because it reaches reasonable achievements even for some of the BERT models, but is still far from the best of them.

The most prominent classifiers in the best models are the Random Forest Regressor, SVR regressor, and Bayesian regressor. They are based on classic machine learning algorithms - Random Forest, Support Vector Machine, and Naive Bayes which are recognized as classic classifiers but strong and good in many ML tasks.

Despite the well-known advantages of preprocessing methods in ML tasks, it seems that there is an overall balance between models that were quicker to preprocess their text and models that worked better on the raw text. It may be that more advanced preprocessing methods such as stemming or lemmatization will be more helpful for learning, but because In most languages it was difficult to find tools that would perform this processing of texts.

6 Conclusions and Future Research

In this paper, we describe our submissions to subtask 1-A and of SemEval-2024.

We applied 17 embedding methods to convert text into vectors, 11 supervised machine learning methods, to predict regression of STR, and did it to 9 different languages.

While our study demonstrates promising outcomes across multiple languages and embedding techniques, a comprehensive error analysis reveals nuanced challenges that warrant further investigation. We observed recurrent patterns of misclassifications, particularly in contexts characterized by linguistic ambiguities, and cultural nuances, and might be affected by the prevalence of sarcasm or irony. These findings highlight the need for robust feature representation and domain-specific adaptations to enhance the accuracy and reliability of sentiment analysis models.

Moreover, our error analysis sheds light on the impact of preprocessing strategies on model performance, revealing a delicate balance between text normalization and the preservation of linguistic subtleties. While preprocessing techniques such as stemming or lemmatization hold promise for improving model generalization, their efficacy varies across languages and datasets, necessitating careful consideration in model development pipelines. We assume that by focusing on one or two languages, we would be able to examine the specific effect of each pre-processing method, as well as focus on the unique characteristics of each language in terms of morphological structure or methods for simplifying and decomposing words, to enable better processing and better results.

There are various ideas for future research regarding the nature of Twitter messages:

(1) use mot preprocessing methods to bring the text to a more understandable shape.

(2) Trying to enrich our training dataset and tune more parameters and longer training because deep learning becomes better with more data to train and more time.

(3) Error analysis must be performed in-depth and repetitive patterns of errors, consistently incorrect classifications, etc. must be identified, to allow for the correction and improvement of the models.

The STR task is an important task that can contribute in many fields, and this study is a milestone in my acquaintance with this task and in developing the way to do it properly.

Language	Classifier	Embedding Type	Pre process	Train Score	Test Score	Rank	SemRel Best Score
Algerian Arabic	RandomForest Regressor	BERT-LaBSE2	No	0.53699	0.44273	17/24	0.68231
Amharic	SVR	BERT-LaBSE2	Yes	0.72871	0.71269	14/18	0.88863
English	SVR	BERT-bert-base-uncased	No	0.75010	0.72020	35/36	0.85958
Hausa	BayesianRidge	BERT-roberta	No	0.61895	0.54304	16/21	0.76429
Kinyarwanda	BayesianRidge	BERT-afrisenti	Yes	0.53506	0.41256	17/21	0.81691
Marathi	SVR	BERT-bert-multi	No	0.76888	0.77817	21/25	0.91086
Moroccan Arabic	SVR	tf-idf	No	0.79914	0.70112	17/23	0.86257
Spanish	BayesianRidge	BERT-robertuito	Yes	0.71538	0.66071	16/25	0.74039
Telugu	MLPRegressor	BERT-distilbert-multi	No	0.74199	0.70555	21/25	0.87336

Table 1: scores of best models for each language in task 1A.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Leo Breiman. 1996. Bagging predictors. Machine learning 24(2), 123-140.

Leo Breiman. 2001. Random forests. Machine learning 45(1), 5-32.

Alexander Budanitsky, and Graeme Hirst. "Evaluating wordnet-based measures of lexical semantic relatedness." Computational linguistics 32.1 (2006): 13-47.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, & Gaël Varoquaux, 2013. API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning (pp. 108–122).

Chih-Chung Chang and Chih-Jen Lin, 2011. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2(3), 1-27.

Sun Chi, Qiu Xipeng, Xu Yige, Huang Xuanjing, 2019. "How to Fine-Tune BERT for Text Classification?." arXiv e-prints: arXiv-1905.

Corinna Cortes and Vladimir Vapnik, 1995. Support-vector networks. Machine learning 20.3 : 273-297.

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.

Bonaventure Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, Chris & Chinenye Emezue, 2022. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. arXiv preprint arXiv:2211.03263.

Evgeniy Gabrilovich and Shaul Markovitch. "Computing semantic relatedness using Wikipedia-based explicit semantic analysis." IJcAI. Vol. 7. 2007.

Yaakov HaCohen-Kerner, Yair Yigal, and Daniel Miller. 2019. The impact of Preprocessing on Classification of Mental Disorders, in Proc. of the 19th Industrial Conference on Data Mining, (ICDM 2019), New York.

Yaakov HaCohen-Kerner, Daniel Miller, and Yair Yigal. 2020. The influence of preprocessing on text classification using a bag-of-words representation, PloS one, vol. 15, p. e0232525.

Taieb Hadj, Mohamed Ali, Torsten Zesch, and Mohamed Ben Aouicha. "A survey of semantic relatedness evaluation datasets and procedures." Artificial intelligence review 53.6 (2020): 4407-4448.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gerard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, & Travis E. Oliphant (2020). Array programming with NumPy. Nature, 585(7825), 357–362.

Ramchoun Hassan, Mohammed Amine Janati Idrissi, Youssef Ghanou, and Mohamed Ettaouil, 2016. "Multilayer Perceptron: Architecture Optimization and Training." International Journal of Interactive Multimedia and Artificial Intelligence 4, no. 1 (2016): 26+.

Tin Kam Ho. 1995. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE.

Ron Keinan and Yaakov HaCohen-Kerner. "JCT at SemEval-2023 Tasks 12 A and 12B: Sentiment Analysis for Tweets Written in Low-resource African Languages using Various Machine Learning and Deep Learning Methods, Resampling, and HyperParameter Tuning." Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). 2023.

HuggingFace API, 2024. https://huggingface.co/docs/api-inference/index Last Access: 13/Feb/2023

Ogueji Kelechi, Yuxin Zhu, and Jimmy Lin, 2021. "Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages." Proceedings of the 1st Workshop on Multilingual Representation Learning.

Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim and Sung Hyon Myaeng, 2006. "Some Effective Techniques for Naive Bayes Text Classification," in IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466, Nov. 2006, doi: 10.1109/TKDE.2006.180.

Jey Han Lau and Baldwin Timothy (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368

Wes McKinney, 2010. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 56 - 61).

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil, 2022. NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 590–602, Marseille, France. European Language Resources Association.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, & Saif M. Mohammad. (2024A). SemRel2024: A Collection of Semantic Textual Relatedness Datasets for 14 Languages.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, & Saif M. Mohammad. (2024B). SemEval-2024 Task 1: Semantic Textual Relatedness. In *"Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)"*.

Juan Ramos (2003). "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. Vol. 242. No. 1.

Guido Van Rossum & Fred Drake, 2009. Python 3 Reference Manual. CreateSpace.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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, & Alexander M. 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 (pp. 38–45). Association for Computational Linguistics.

Appendix A - Details of the Data Sets

Language	Train Size	Dev Size	Test Size
Algerian Arabic	949	97	583
Amharic	599	95	171
English	911	249	919
Hausa	558	212	565
Kinyarwanda	435	102	222
Marathi	270	267	284
Moroccan Arabic	319	70	324
Spanish	615	139	599
Telugu	260	130	273

Appendix B – All Embedding Models

Model Name	Type	Languages
doc2vec	basic	Multilingual
mUSE		
tf-idf		
LSA		
LDA		
distiluse-base-multilingual-cased-v2	Sentence-Transformers	
paraphrase-multilingual-MiniLM-L12-v2		
LaBSE		
clip-ViT-B-32-multilingual-v1		
bert-base-multilingual-uncased	BERT	
lxyuan/distilbert-base-multilingual-cased-sentiments-student		
Davlan/afrisenti-twitter-sentiment-afroxlmr-large		
intfloat/multilingual-e5-base		
l3cube-pune/indic-sentence-similarity-sbert		
setu4993/LaBSE		
setu4993/LEALLA-large		
FacebookAI/xlm-roberta-base		
Abdou/arabert-large-algerian		Algerian + Moroccan
alger-ia/dziribert		
CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment		
asafaya/bert-large-arabic		
aubmindlab/bert-base-arabert		
SI2M-Lab/DarijaBERT		
pysentimiento/robertuito-sentiment-analysis		Spanish
llange/xlm-roberta-large-spanish		
dccuchile/bert-base-spanish-wwm-uncased		
maxpe/bertin-roberta-base-spanish_sem_eval_2018_task_1		
cardiffnlp/twitter-roberta-base-sentiment-latest		English
distilbert-base-uncased-finetuned-sst-2-english		
bert-base-uncased		
roberta-large		

Appendix C - Full Results, 10 Best Models For Every Language

Language	Classifier	Embedding Type	Preprocessing	Train_Score
Algerian Arabic	RandomForestRegressor	BERT-LaBSE2	No	0.5369947229
Algerian Arabic	GradientBoostingRegressor	BERT-LaBSE2	No	0.5292887473
Algerian Arabic	RandomForestRegressor	BERT-LaBSE2	Yes	0.5253867143
Algerian Arabic	SVR	BERT-bert-multi	Yes	0.5220256197
Algerian Arabic	SVR	BERT-bert-multi	No	0.5210616209
Algerian Arabic	BayesianRidge	BERT-aubmindlab	No	0.5152289012
Algerian Arabic	BayesianRidge	BERT-aubmindlab	Yes	0.5137481244
Algerian Arabic	SVR	SenTransformers-LaBSE	No	0.5110653241
Algerian Arabic	SVR	SenTransformers-LaBSE	Yes	0.5104857707
Algerian Arabic	SVR	BERT-LaBSE2	No	0.5077027734
Amharic	SVR	BERT-LaBSE2	Yes	0.7287084049
Amharic	BayesianRidge	BERT-roberta	Yes	0.7246094157
Amharic	BayesianRidge	BERT-roberta	No	0.7204889719
Amharic	MLPRegressor	BERT-roberta	Yes	0.7080872218
Amharic	BayesianRidge	BERT-LaBSE2	Yes	0.7044388055
Amharic	SVR	BERT-LaBSE2	No	0.7023932501
Amharic	MLPRegressor	BERT-roberta	No	0.6991415451
Amharic	BayesianRidge	BERT-LaBSE2	No	0.694283367
Amharic	BayesianRidge	SenTransformers-LaBSE	Yes	0.6608416741
Amharic	Ridge	SenTransformers-LaBSE	Yes	0.6608308762
English	SVR	BERT-bert-multi	No	0.7582006981
English	SVR	BERT-bert-base-uncased	No	0.750103082
English	BayesianRidge	BERT-bert-roberta-large	No	0.741107994
English	SVR	BERT-LaBSE2	No	0.7404424659
English	BayesianRidge	BERT-bert-multi	No	0.735515388
English	Ridge	BERT-bert-roberta-large	No	0.7322067128
English	BayesianRidge	BERT-bert-roberta-large	Yes	0.731029189
English	SVR	BERT-LaBSE2	Yes	0.7307237556
English	Ridge	BERT-bert-roberta-large	Yes	0.7270267272
English	SVR	BERT-twitter-roberta	No	0.7231381043
Hausa	BayesianRidge	BERT-roberta	No	0.6189488918
Hausa	MLPRegressor	BERT-roberta	Yes	0.6104493667
Hausa	BayesianRidge	BERT-roberta	Yes	0.6077610956
Hausa	MLPRegressor	BERT-roberta	No	0.6028932623
Hausa	SVR	BERT-afrisenti	No	0.5847085719
Hausa	SVR	BERT-afrisenti	Yes	0.5814811298
Hausa	SVR	BERT-LaBSE2	Yes	0.5483401586
Hausa	BayesianRidge	BERT-afrisenti	Yes	0.5479463345
Hausa	SVR	BERT-bert-multi	No	0.5371360093
Hausa	BayesianRidge	BERT-afrisenti	No	0.5369844352

Kinyarwanda	BayesianRidge	BERT-afrisenti	Yes	0.5350585294
Kinyarwanda	SVR	BERT-afrisenti	Yes	0.5146730111
Kinyarwanda	SVR	BERT-e5-base	No	0.5136255749
Kinyarwanda	BayesianRidge	BERT-distilbert-multi	No	0.505681762
Kinyarwanda	BayesianRidge	BERT-e5-base	Yes	0.4963074713
Kinyarwanda	SGDRegressor	BERT-e5-base	Yes	0.4960261965
Kinyarwanda	MLPRegressor	BERT-roberta	No	0.4956656724
Kinyarwanda	BayesianRidge	BERT-e5-base	No	0.4947476356
Kinyarwanda	SGDRegressor	BERT-e5-base	No	0.4933142053
Kinyarwanda	GradientBoostingRegressor	BERT-distilbert-multi	No	0.4911255779
Marathi	SVR	BERT-bert-multi	No	0.768881107
Marathi	BayesianRidge	BERT-bert-multi	No	0.7688210054
Marathi	SVR	BERT-bert-multi	Yes	0.7546816577
Marathi	BayesianRidge	BERT-distilbert-multi	No	0.7532801443
Marathi	BayesianRidge	BERT-bert-multi	Yes	0.7505721435
Marathi	SVR	BERT-distilbert-multi	No	0.7467252478
Marathi	MLPRegressor	BERT-bert-multi	No	0.7440670356
Marathi	BayesianRidge	BERT-distilbert-multi	Yes	0.7415936477
Marathi	SVR	BERT-distilbert-multi	Yes	0.7414378556
Marathi	MLPRegressor	BERT-distilbert-multi	No	0.7379256945
Moroccan Arabic	SVR	tf-idf	No	0.7991443722
Moroccan Arabic	SVR	tf-idf	Yes	0.796339094
Moroccan Arabic	Ridge	SenTransformers-LaBSE	Yes	0.7787889425
Moroccan Arabic	SVR	BERT-LaBSE2	No	0.7778968174
Moroccan Arabic	BayesianRidge	SenTransformers-LaBSE	Yes	0.777541694
Moroccan Arabic	MLPRegressor	SenTransformers-LaBSE	Yes	0.772735299
Moroccan Arabic	SVR	BERT-LaBSE2	Yes	0.77245585
Moroccan Arabic	SVR	SenTransformers-LaBSE	Yes	0.7704490304
Moroccan Arabic	BayesianRidge	BERT-LaBSE2	No	0.7676106274
Moroccan Arabic	BayesianRidge	BERT-CAMeL-Lab	No	0.7665164306
Spanish	BayesianRidge	BERT-robertuito	Yes	0.7153770062
Spanish	GradientBoostingRegressor	BERT-robertuito	Yes	0.7128803162
Spanish	BayesianRidge	BERT-robertuito	No	0.7112746809
Spanish	AdaBoostRegressor	BERT-robertuito	Yes	0.6989013087
Spanish	BayesianRidge	BERT-bert-base-spanish	Yes	0.6979171296
Spanish	SGDRegressor	BERT-distilbert-multi	No	0.6971499681
Spanish	GradientBoostingRegressor	BERT-robertuito	No	0.6967140825
Spanish	BayesianRidge	BERT-distilbert-multi	Yes	0.6964474195
Spanish	SVR	BERT-LaBSE2	Yes	0.6959682585
Spanish	SGDRegressor	BERT-distilbert-multi	Yes	0.6956928668
Telugu	MLPRegressor	BERT-distilbert-multi	No	0.7419850235
Telugu	SVR	BERT-LaBSE2	Yes	0.7331294075

Telugu	SVR	BERT-bert-multi	No	0.7325062071
Telugu	BayesianRidge	BERT-distilbert-multi	No	0.7313601112
Telugu	SVR	BERT-distilbert-multi	No	0.7312296203
Telugu	BayesianRidge	BERT-bert-multi	No	0.7255846168
Telugu	SVR	BERT-bert-multi	Yes	0.722301082
Telugu	SGDRegressor	BERT-distilbert-multi	No	0.7199655333
Telugu	BayesianRidge	BERT-distilbert-multi	Yes	0.7196172815
Telugu	SVR	BERT-LaBSE2	No	0.7194733505