electronics

MDPI

Article

Combining Transformer Embeddings with Linguistic Features for Complex Word Identification

Jenny A. Ortiz-Zambrano [1,*,†], César Espin-Riofrio [1,†] and Arturo Montejo-Ráez [2,*,†]

[1] Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Guayaquil 090514, Ecuador
[2] Departamento de Informática, Universidad de Jaén, 23071 Jaén, Spain
[*] Correspondence: jenny.ortizz@ug.edu.ec (J.A.O.-Z.); amontejo@ujaen.es (A.M.-R.)
[†] These authors contributed equally to this work.

Abstract: Identifying which words present in a text may be difficult to understand by common readers is a well-known subtask in text complexity analysis. The advent of deep language models has also established the new state-of-the-art in this task by means of end-to-end semi-supervised (pre-trained) and downstream training of, mainly, transformer-based neural networks. Nevertheless, the usefulness of traditional linguistic features in combination with neural encodings is worth exploring, as the computational cost needed for training and running such networks is becoming more and more relevant with energy-saving constraints. This study explores lexical complexity prediction (LCP) by combining pre-trained and adjusted transformer networks with different types of traditional linguistic features. We apply these features over classical machine learning classifiers. Our best results are obtained by applying Support Vector Machines on an English corpus in an LCP task solved as a regression problem. The results show that linguistic features can be useful in LCP tasks and may improve the performance of deep learning systems.

Keywords: lexical complexity prediction; linguistic features; features fusion; pre-trained large language models

check for updates

Citation: Ortiz-Zambrano, J.A.; Espin-Riofrio, C.; Montejo-Ráez, A. Combining Transformer Embeddings with Linguistic Features for Complex Word Identification. *Electronics* **2023**, *12*, 120. https://doi.org/10.3390/electronics12010120

Academic Editor: Min Zhang

Received: 7 November 2022
Revised: 9 December 2022
Accepted: 22 December 2022
Published: 27 December 2022

1. Introduction

A general assumption is that those who are familiar with the vocabulary of a text are often able to understand it even if there is a problem with the grammatical structure. The task of recognizing words in document content that is difficult or complex for a particular group of people is called complex word identification (CWI) [1] and is the basis of many related applications where text simplification is involved. Automatic lexical simplification can be an effective way to make text accessible to different audiences [2]. A complex word is considered one that is difficult to understand by a reader with an average level of literacy. In a more general view [3], lexical complexity prediction (LCP) tries to assign a score of complexity to values, turning the task into a regression problem instead of a binary classification task.

Deep learning and its innovative technologies represent cutting-edge new technologies for various natural language processing (NLP) tasks [4]. LCP is not an exception [5]. After comparing and analysing deep learning approaches with classic approaches, possible solutions are feasible for English and resource-poor deep learning languages, where deep models are not always available or functional. It should also be noted that the computational requirements for applying deep learning models are significantly higher than those of traditional approaches [6].

The field of NLP has made tremendous progress in the last years, especially thanks to the Transformer architecture [7]. This architecture uses a large amount of untagged text corpus [8] to train the network. Deep learning models are significantly improved over "shallow" machine learning models with the advent of transfer learning and pre-trained language models. The BERT and XLM-RoBERTa pre-trained deep learning language models are considered to be at the forefront of many NLP tasks [9].

This research focuses on evaluation of the improvement of the results of the prediction of complex words aimed at the English language by applying technological solutions with an emphasis on Deep Learning models.

Our approach takes advantage of the combination of classic NLP techniques and their integration in Transformer-based deep learning models: BERT [10] and XLM-RoBERTa [11]. These techniques provide a set of characteristics of a different nature: linguistic, syntactic, statistical, and semantic. The experiments are carried out with the English CompLex 2.0 corpus described in [3].

Our challenge is to improve the prediction of the lexical complexity task presented in the SemEval 2021 competition, implementing a refined model executed on a previously pre-trained model for which we have followed several of the recommendations applied by the different winning competition teams [12].

For our evaluation, we apply various metrics to the regression algorithms. The measurement is made based on the Mean Absolute Error. Our best results achieve very significant performance: showing a competitive MAE of 0.0688 and a Pearson score of 0.8911 for the identification of simple vs. complex words.

2. Related Work

For the past few decades, calculation of the number of syllables in a word [13], checking if a word is part of a particular list, and, accordingly, classifying it as simple or complex were among the earliest approaches [14]. Across the years, some machine-learning systems appeared and categorised words based on their characteristics. Breiman [15] used contextual, lexical, and semantic characteristics with the application of the random forest classifier to determine if a word was complex. In this system, a total of 45 linguistic features were calculated, and each word was modelled as a feature vector. The system applied surface features, dependency tree features, corpus-based features, and WordNet-related features. The best results achieved had an accuracy of 0.186, a recall of 0.673, a G score of 0.750, and an F score of 0.292.

Most of the research on text complexity over the past few years has focused on complex word identification (CWI). The goal of these applications is to reduce word complexity based on the composition of different indicative features, as outlined in the work done by Shardlow et al. [3], which presents a new data set for LCP, the CompLex corpus, and an approach to a set of features in the word embedding of Gloves, InferSent, and various language features obtained as predictive sources of vocabulary complexity, such as word frequency, word length, or number of syllables. Then, they trained a linear regression model using different subsets of functions, obtaining an MAE value of 0.0853.

Lately, Shardlow et al. [12] developed a word complexity prediction system for common LCP tasks hosted on the SemEval 2021 evaluation campaign. The task organizers distributed an updated version of the CompLex corpus to participants. The task was in the lexical semantics track, which consisted of predicting the value of word complexity in context.

Ortiz-Zambrano and Montejo-Ráez [16] designed a machine learning approach based on word level and 15 language features. They trained a supervised random forest regression algorithm for a set of features. Several runs were performed with different values to observe the performance of the algorithm. The best results obtained an MAE of 0.07347, MSE equal to 0.00938, and an RMSE of 0.096871.

El Mamoun et al. [17] introduced a new deep learning-based system for this challenging task. The proposed system consisted of a deep learning model based on a pre-trained transformer encoder for word and Multi-Word Expression (MWE) complexity prediction. First, on top of the encoder's contextualized word embedding, the model employs an attention layer on the input context and the complex word or MWE. They investigated both single-task and joint training on both sub-tasks' data using multiple pre-trained transformer-based encoders. The obtained results are very promising and show the effectiveness of fine-tuning pre-trained transformers for LCP.

Uluslu [2] presented the first automatic lexical simplification system for the Turkish language. They presented a new text simplification pipeline based on a pre-trained representation BERT model together with morphological features to generate grammatically correct and semantically appropriate word-level simplifications.

3. Dataset

We carried out the experiments by applying the augmented version of the corpus CompLex proposed by [3]. It is the first multi-domain data set in English where selected words are labelled as difficult with a level of complexity on a five-point Likert scale: Very Easy (0), Easy (0.25), Neutral (0.5), Difficult (0.75), and Very Difficult (1.0). The annotation process was done manually using a crowd-working service, collecting a total of 20 annotations per entry by native English speakers. The final complexity score was calculated by summing up all the Likert points from the annotators and dividing by 100. Complex words from three sources/domains were labelled: the Bible, Europarl, and biomedical texts. Some sample entries from the data set are shown in Table 1.

Table 1. Sample entries from CompLex.

Id	Corpus	Sentence	Token	Complexity
3ZL....32A	Bible	Behold, there came up out of the river seven cattle, sleek and fat, and they fed in the marsh grass.	river	0.100
34R...E5C	Bible	I am a fellow bondservant with you and with your brothers, the prophets, and with those who keep the words of this book.	brothers	0.400
3GM...UY3	Biomed	Supplementary data are available at NAR online.	online	0.107
3KI...67D	Biomed	In lens epithelium derived from alphaAKO lenses, cell growth rates were reported to be 50% lower compared to wild type, suggesting a role for alphaA in regulating the cell cycle.	growth	0.107
3VA...PSC	Europarl	(ES) Mr President, as a Spanish Member resident in the Canary Islands, I want to thank you for remembering the victims of the accident on 20 August.	victims	0.191
3H6...PWP	Europarl	Over 40% of the energy we use is consumed in buildings and 75% of the buildings standing today will still be here in 2050, so we need to tackle energy efficiency in existing buildings as well as in new stock.	efficiency	0.3333

This resource was delivered as a reference collection for the 15th International Workshop on Semantic Evaluation in Task 1: Lexical Complexity Prediction. Table 2 presents the statistics for CompLex 2.0. The SemEval 2021 Task 1 consisted of two subtasks: Subtask 1 was based on single words, and Subtask 2 focused on multi-word expressions (MWE). The competition interested 198 teams in total, of which 54 teams presented official predictions in the test data for Subtask 1 and 37 for Subtask 2 [12].

Table 2. The statistics for CompLex 2.0.

Subset	Genre	Context	Unique Tokens	Average Complexity
All	**Total**	**10,800**	**5617**	**0.321**
	Europarl	3600	2227	0.303
	Biomed	3600	1904	0.353
	Bible	3600	1934	0.307
Single	**Total**	**9000**	**4129**	**0.302**
	Europarl	3000	1725	0.286
	Biomed	3000	1388	0.325
	Bible	3000	1462	0.293
MWE	**Total**	**1800**	**1488**	**0.419**
	Europarl	600	502	0.388
	Biomed	600	516	0.491
	Bible	600	472	0.377

4. System Description

Our approach is based on the combination of encodings from pre-trained language models with different linguistic features to produce a final vector that is fed into some classic machine learning algorithms, as is graphically described in Figure 1. As can be seen, from a given text, two feature vectors are generated: the embeddings from the transformer model and the linguistic features computed from the text (normalized with a Z-score transformation). These vectors are combined and serve as input to the classification algorithm.

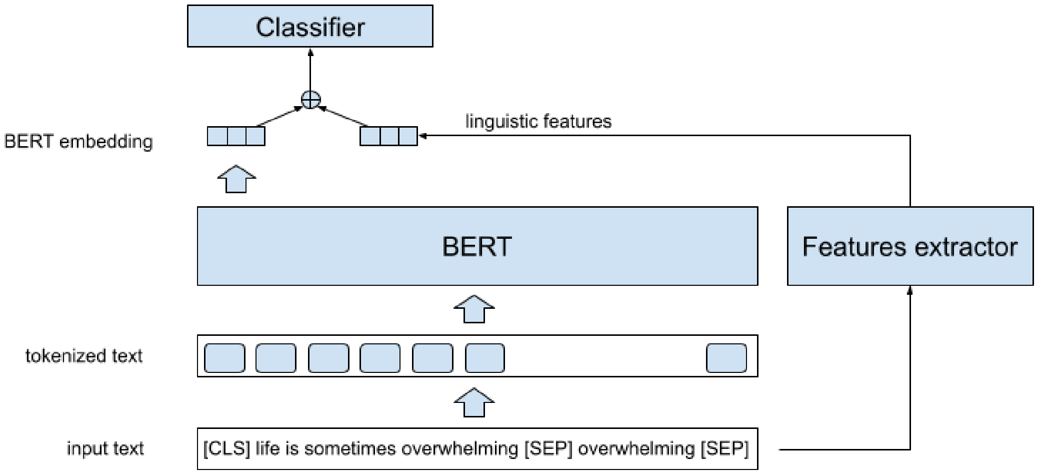

Figure 1. Architecture of the system.

As research in recent times has shown that Transformers achieve significant performance for most NLP tasks [18], we take the initiative applied by the top-ranking teams in the SemEval task and select these two well-known large language models:

- BERT represents the baseline of Transformer-based models. This model has been extensively pre-trained on English [10].
- RoBERTa usually performs better on downstream tasks, as it improves upon BERT by modifying key hyperparameters and trains with mini-batches and higher learning rates [19].

The features considered in our research are described in the next section. Several machine learning algorithms for training the final classifier have been explored. They are detailed in Section 4.2.

The BERT and RoBERTa models are fine-tuned with final dense layers as the classifier network using the architecture used by default in the sequence classification version of these models, as it is the Huggingface implementation of these architectures. No hyperparameter exploration is done, so, again, the default parametrization from the taken implementation is applied. Regarding classic machine learning methods, they are trained with the default parameters as set by Scikit-learn implementations.

4.1. Features

We compute some morphological aspects of the text (23 linguistic features). Then, several experiments are performed by combining the 23 linguistic features with the word and sentence embeddings from pre-trained and fine-tuned deep learning models, as in the work done by Liebeskind et al. [20]. Therefore, the embeddings at the sentence level from which the token comes and the embeddings at the word level are obtained, as the information from the context of the token constitutes important help so that the embeddings at the word level can be adequate for LCP [21].

The linguistic features are computed as classic features in computational linguistics [20,22], and the "semantic features" are those that result from the word- and sentence-level embeddings described previously. Linguistic features have proven to be a great contribution in the prediction of lexical complexity; therefore, we rely on the features applied by [16]. Finally, all these features are normalized by applying a z-score transformation to be later executed by the different supervised learning algorithms.

The properties contemplated in the linguistic features include two families of properties: morphological and syntactic. Specifically, eight of the linguistic features contain information extracted from the POS tagger [20]. We apply the *en_core_news_sm* model from Spacy's statistical POS tagger https://spacy.io/ (accessed 5 August 2022) to extract them. We use the following eight tags from the Universal POS tags: PROPN, AUX, VERB, ADP, NOUN, NN, SYM, and NUM [20,23–27]:

1. *PROPN*: Number of pronouns within the sentence.
2. *AUX*: Number of auxiliaries within the sentence.
3. *VERB*: Number of verbs within the sentence.
4. *ADP*: Number of adverbs within the sentence.
5. *NOUN*: Number of nouns within the sentence.
6. *NN*: Number of nouns, singular or massive.
7. *SYM*: Number of symbols within the sentence.
8. *NUM*: Number of numbers within the sentence.
9. *Absolute frequency*: the absolute frequency.
   The frequency of words is a measure that serves as an indicator of lexical complexity. If, in common parlance, a word occurs frequently, it is more likely to be recognized [3,28].
10. *Relative frequency*: the relative frequency of the target word.
11. *Word length*: the number of characters of the token. The length of the word is calculated as the number of its characters. It is often the case that longer words are more difficult to process and can therefore be considered *complex* [22,24,29].
12. *Number of syllables*: the number of syllables. A good estimate of complexity is the number of syllables contained in a word [3,24,25,29].
13. *Target word position* (token-position): the position of the target word in the sentence. Position of the word (WordPosition) [25,29].
14. *Number of words in the sentence*: number of words in the sentence. Words in sentence (NumSentenceWords) [25,29].
    Based on the work proposed by [25] for exploring linguistic features for lexical complexity prediction, we implemented:
15. *Part of Speech (POS)*: the Part of Speech category.
16. *Relative frequency of the previous token*: the relative frequency of the word before the token.

17. *Relative frequency of the word after the token*: the relative frequency of the word after the token.
18. *Length of the previous word*: the number of characters in the word before the token.
19. *Length of the following word*: the number of characters in the word after the token.
20. *Lexical diversity—MTDL*: the lexical diversity of the target word in the sentence. Additionally, the following WordNet features were also considered for each target word, as in the work carried out by [26]:
21. *Number of synonyms* [22].
22. *Number of hyponyms* [22].
23. *Number of hyperonyms* [22].

Table 3 contains the description of the abbreviations for a better understanding of the features used in our experiments.

Table 3. Description of the explored feature sets.

Feature Identifier	Description
LF	Linguistic Features.
BERT$_{sent}$	Sentence encodings from BERT model.
BERT$_{word}$	Token/word encodings from BERT model.
XLMR$_{sent}$	Sentence encodings from XLM-RoBERTa model.
XLMR$_{word}$	Token/word encodings from RoBERTa model.

4.2. Traditional Machine Learning Classifiers

With the purpose of achieving the highest performance of the algorithms in terms of the prediction of the lexical complexity of the words, a total of eight supervised algorithms for regression were applied for the training and evaluation of the different data sets. These are:

1. AdaBoost—AB [30].
2. Decision Tree—DT [31].
3. Gradient Boost—GB [23].
4. Stochastic Gradient—SG [32].
5. Nearest Neighbors—KNN [20].
6. Support Vector Machines—SVM [20].
7. Passive Aggressive—PA [33].
8. Random Forest—RF [21,27].

5. Results

First, we obtain the morphological aspects of the text; we perform several experiments applying the 23 linguistic features and combine them with the word and sentence embeddings of pre-trained and fine-tuned deep learning models.

We explored both pre-trained and fine-tuned BERT and RoBERTa models with the coded token and sentence. We employ multiple training strategies for pre-trained models or fine-tuned ones, with or without linguistic features, taking the embedding of the target word or that of the full sentence, etc. The Table 4 presents the results achieved after the different executions.

Table 4. Results for different feature combinations, models, and classifiers.

Results for the Deep Learning Approaches with CompLex							
Configuration	**Model**	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**	**R2**
BERT-W ⊕ BERT-S ⊕ LF	fine-tuned	**SVR**	**0.068898**	**0.009908**	**0.094294**	**0.891137**	**0.80**
BERT-W⊕ BERT-S ⊕ XLMR-W⊕ XLMR-S⊕ LF	*fine-tuned*	SVR	0.068899	0.009908	0.094296	0.8911367	0.80
BERT-W⊕ BERT-S⊕ XLMR-W⊕ XLMR-S⊕ LF	*pre-trained*	SVR	0.068899	0.009908	0.094296	0.891136	0.79

Table 4. *Cont.*

Results for the Deep Learning Approaches with CompLex							
Configuration	**Model**	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**	**R2**
BERT-W⊕ BERT-S⊕ LF	*fine-tuned*	GBR	0.068972	0.011898	0.099297	0.927208	0.87
BERT-W⊕ BERT-S⊕ XLMR-W⊕ XLMR-S⊕ LF	*fine-tuned*	GBR	0.068974	0.011898	0.099297	0.927218	0.86
BERT-W⊕ BERT-S⊕ XLMR-W⊕ XLMR-S⊕ LF	*pre-trained*	GBR	0.068974	0.011898	0.099297	0.927218	0.87
BERT-W⊕ LF	*fine-tuned*	SVR	0.069900	0.009913	0.094392	0.890911	0.79
BERT-W⊕ LF	*fine-tuned*	GBR	0.070124	0.012018	0.099372	0.926726	0.86
BERT-W⊕ BERT-S	*fine-tuned*	SVR	0.071623	0.009204	0.095901	0.874019	0.77
BERT-W⊕ LF	*pre-trained*	SVR	0.074342	0.009909	0.095558	0.864029	0.75
BERT-W⊕ BERT-S	*pre-trained*	SVR	0.074394	0.009323	0.096034	0.873943	0.75
BERT-W⊕ LF	*pre-trained*	GBR	0.075430	0.012348	0.100003	0.900043	0.80
BERT-W	*pre-trained*	RLM	0.075552	0.009734	0.098232	0.789901	0.63
BERT-W	*fine-tuned*	SVR	0.075897	0.009520	0.097119	0.864816	0.76
BERT-W	*pre-trained*	SVR	0.075938	0.009559	0.097123	0.864	0.76

We evaluate the results of our system by applying the regression metrics in the execution of the supervised learning algorithms, specifically MAE, MSE, RMSE, and R2. We emphasize that we apply the methodologies of the winning teams, which are based on the application of language models based on the pre-trained and adjusted Transformers BERT and RoBERTa [9,34,35], together with the linguistic, syntactic and statistical characteristics, and the embedding results at the word and sentence level.

The results obtained are: MAE of 0.0688, MSE of 0.0099, R2 of 0.80, and Pearson of 0.8911 with execution of the Super Vector Regressor regression algorithm; the results can be seen in Table 4.

The Table 4 presents the 15 best results. Only the SVR, GBR, and RFR algorithms appear in the first places. The Support Vector Regressor (SVR) algorithm is the one with the best performance for predicting the complexity of simple words, followed by the Gradient Boosting (GB) algorithm, which achieves very interesting results, and finally the Random Forest Regressor (RFR).

6. Discussion

The Table 5 presents the results of Subtask 1 in SemEVal 2021. Results are ranked in terms of Pearson, MAE, MSE, and R^2. The main rank order corresponds to MAE.

Table 5. Subtask 1: results and rank in terms of Pearson, MAE, MSE, and R^2 in SemEval 2021. The rank evaluation corresponds to MAE.

Rank	Team Name	Pearson	MAE	MSE	R^2
1	JUST Blue	0.7886	0.0609	0.0062	0.6172
2	DeepBlueAI	0.7882	0.0610	0.0061	0.6210
3	OCHADAI-KYOTO	0.7772	0.0617	0.0065	0.6015
4	ia pucp	0.7704	0.0618	0.0066	0.5929
5	Alejandro M.	0.7790	0.0619	0.0064	0.6062
	Our system	-	0.0875	0.0131	0.1930

Based on the results achieved by the participants, Table 5 shows that the winning team achieves an MAE of 0.0609, an MSE of 0.0062, R2 of 0.6172, and a Pearson of 0.7886.

The total of 23 linguistic features are fed into a Random Forest Regressor algorithm. We train the algorithm with the evaluation data and predict the results of the test set with the trained model. The results obtained with a focus on the prediction of simple words are: MAE of 0.0875, MSE of 0.0131, and R2 of 0.1930. Considering the number of competitors is quite large (54 teams) and that the result presented by the first place winner is an MAE of 0.0609, we see that there is a small difference, which allows us to trust our simple approach.

Several of the teams ran the data sets with the BERT and RoBERTa models, such as JUST BLUE, RG PA, Andi, CS-UM6P, and OCHADAI-KYOTO, just to name a few [12].

It should be noted that linguistic features play an important role in predicting word complexity, and they attract applications from many researchers who include them as part of their data sets.

7. Conclusions

In this research, we presented a contribution to the prediction of the complexity of simple words in the English language based on the execution of several language models based on BERT and XLM-RoBERTa models in combination with different linguistic features.

Multiple experiments were carried out using various regression algorithms and their configuration to obtain the maximum performance for the different data sets. Several models based on pre-trained and adjusted Transformers were applied on the different English data sets. The embeddings obtained with the execution of the adjusted models in conjunction with the manual features achieved better performance with the execution of the machine learning algorithms explored, which led to competitive results with the Support Vector Regressor (SVR) algorithm with a fine-tuned BERT model.

As a possible alternative proposal to achieve better prediction of lexical complexity, we are very interested in further experimentation on data sets for the English language and other languages where resources in the area of Lexical Simplification are scarce, such as, for example, the Spanish language, testing the latest generation of Transformer models. For this, the extrinsic evaluation will be overcome to compare the best systems for this specific task with the possibilities of integrating external features such as those proposed in this work.

The BERT and RoBERTa Transformer-based language models were applied to a set of features from a training corpus of simple English words. Our solution is not geared towards multi-words. We see it convenient to apply other language models that allow us to evaluate the results as a possible improvement in the level of word prediction. The value of R2 obtained in our experiments tells us that only 19% of the variation of the complexity can be explained from our features, which is minimal. This suggests that further experimentation must be done with additional features and ablation tests to identify which kinds of features (embeddings, complexity related, stylistically related, etc.) contribute more to the identification of difficult words.

The experimentation made and the proposed system for lexical simplification can help other researchers in their work to explore such a design. Pre-trained models and classical algorithms do not demand high computing resources and could be a solution in low-resourced environments, such as mobile devices and the like.

Author Contributions: Conceptualization and methodology, J.A.O.-Z. and A.M.-R.; software and validation, J.A.O.-Z. and C.E.-R.; data preparation and experimentation, J.A.O.-Z.; analysis, J.A.O.-Z. and A.M.-R., original draft preparation, J.A.O.-Z.; review and editing, C.E.-R. and A.M.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially funded by the WeLee Project grant 1380939 (FEDER Andalucía 2014-2020) from the Andalusian Regional Government of Spain.

Data Availability Statement: The CompLex data set is available at https://github.com/MMU-TDMLab/CompLex, accessed on 6 November 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rico-Sulayes, A. General lexicon-based complex word identification extended with stem n-grams and morphological engines. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain, 23 September 2020.
2. Uluslu, A.Y. Automatic Lexical Simplification for Turkish. *arXiv* **2022**, arXiv:2201.05878.

3. Shardlow, M.; Cooper, M.; Zampieri, M. CompLex: A New Corpus for Lexical Complexity Predicition from Likert Scale Data. In Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), Marseille, France, 11 May 2020.
4. Singh, S.; Mahmood, A. The NLP cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access* **2021**, *9*, 68675–68702. [CrossRef]
5. Nandy, A.; Adak, S.; Halder, T.; Pokala, S.M. cs60075_team2 at SemEval-2021 Task 1: Lexical Complexity Prediction using Transformer-based Language Models pre-trained on various text corpora. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, 5–6 August 2021; pp. 678–682.
6. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Virtual Event, 3–10 March 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623. [CrossRef]
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems NIPS 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
8. Canete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish pre-trained BERT model and evaluation data. In Proceedings of the PML4DC, ICLR 2020, Addis Ababa, Ethiopia, 26 April–1 May 2020.
9. Yaseen, T.B.; Ismail, Q.; Al-Omari, S.; Al-Sobh, E.; Abdullah, M. JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; pp. 661–666.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
11. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8440–8451. [CrossRef]
12. Shardlow, M.; Evans, R.; Paetzold, G.H.; Zampieri, M. SemEval-2021 Task 1: Lexical Complexity Prediction. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, 25 May 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1–16. [CrossRef]
13. Mc Laughlin, G.H. SMOG grading-a new readability formula. *J. Read.* **1969**, *12*, 639–646.
14. Dale, E.; Chall, J.S. A formula for predicting readability: Instructions. *Educ. Res. Bull.* **1948**, *27*, 37–54.
15. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
16. Ortiz-Zambrano, J.A.; Montejo-Ráez, A. Complex words identification using word-level features for SemEval-2020 Task 1. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, 25 May 2021; pp. 126–129.
17. El Mamoun, N.; El Mahdaouy, A.; El Mekki, A.; Essefar, K.; Berrada, I. CS-UM6P at SemEval-2021 Task 1: A Deep Learning Model-based Pre-trained Transformer Encoder for Lexical Complexity. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, 25 May 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 585–589. [CrossRef]
18. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
19. Zhuang, L.; Wayne, L.; Ya, S.; Jun, Z. A Robustly Optimized BERT Pre-training Approach with Post-training. In Proceedings of the 20th Chinese National Conference on Computational Linguistics, Huhhot, China, 13–15 August 2021; Chinese Information Processing Society of China: Huhhot, China, 2021; pp. 1218–1227.
20. Liebeskind, C.; Elkayam, O.; Liebeskind, S. JCT at SemEval-2021 Task 1: Context-aware Representation for Lexical Complexity Prediction. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; pp. 138–143.
21. Zaharia, G.E.; Cercel, D.C.; Dascalu, M. UPB at SemEval-2021 Task 1: Combining Deep Learning and Hand-Crafted Features for Lexical Complexity Prediction. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) Bangkok, Thailand, 5–6 August 2021; Association for Computational Linguistics: Minneapolis, MN, USA, 2021; pp. 609–616. [CrossRef]
22. Mosquera, A. Alejandro Mosquera at SemEval-2021 Task 1: Exploring Sentence and Word Features for Lexical Complexity Prediction. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; Association for Computational Linguistics: Minneapolis, MN, USA, 2021; pp. 554–559. [CrossRef]
23. Vettigli, G.; Sorgente, A. CompNA at SemEval-2021 Task 1: Prediction of lexical complexity analyzing heterogeneous features. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; pp. 560–564.

24. Paetzold, G.; Specia, L. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; pp. 969–974.
25. Ronzano, F.; Anke, L.E.; Saggion, H. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; pp. 1011–1016.
26. Gooding, S.; Kochmar, E. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, LA, USA, 5–6 June 2018; pp. 184–194.
27. Desai, A.T.; North, K.; Zampieri, M.; Homan, C. LCP-RIT at SemEval-2021 Task 1: Exploring Linguistic Features for Lexical Complexity Prediction. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; Association for Computational Linguistics: Minneapolis, MN, USA, 2021; pp. 548–553. [CrossRef]
28. Rayner, K.; Duffy, S.A. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Mem. Cogn.* **1986**, *14*, 191–201. [CrossRef] [PubMed]
29. Shardlow, M. A Comparison of Techniques to Automatically Identify Complex Words. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop, Sofia, Bulgaria, 5–7 August 2013; pp. 103–109.
30. Paetzold, G. UTFPR at SemEval-2021 Task 1: Complexity Prediction by Combining BERT Vectors and Classic Features. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; pp. 617–622.
31. Shardlow, M.; Evans, R.; Zampieri, M. Predicting lexical complexity in English texts: the Complex 2.0 dataset. *Lang. Resour. Eval.* **2022**, *56*, 1153–1194. [CrossRef]
32. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the COMPSTAT'2010, 19th International Conference on Computational Statistics, Paris, France, 22–27 August 2010; pp. 177–186.
33. Crammer, K.; Dekel, O.; Keshet, J.; Shalev-Shwartz, S.; Singer, Y. Online Passive-Aggressive Algorithms. *J. Mach. Learn. Res.* **2006**, *7*, 551–585.
34. Song, B.; Pan, C.; Wang, S.; Luo, Z. DeepBlueAI at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with Stacking Diverse Language Model-Based Methods. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; pp. 1130–1134.
35. Taya, Y.; Kanashiro Pereira, L.; Cheng, F.; Kobayashi, I. OCHADAI-KYOTO at SemEval-2021 Task 1: Enhancing Model Generalization and Robustness for Lexical Complexity Prediction. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; Association for Computational Linguistics: Minneapolis, MN, USA, 2021; pp. 17–23. [CrossRef]

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