UMUTeam and SINAI at SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis using Multilingual Large Language Models and Data Augmentation

José Antonio García-Díaz[1], **Ronghao Pan**[1], **Salud María Jiménez Zafra**[2], **María-Teresa Martín-Valdivia**[2], **L. Alfonso Ureña-López**[2], **Rafael Valencia-García**[1]

[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{joseantonio.garcia8, ronghao.pan, valencia}@um.es

[2]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain
{sjzafra, maite, laurena}@ujaen.es

Abstract

This work presents the participation of the UMUTeam and the SINAI research groups in the SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis. The goal of this task is to predict the intimacy of a set of tweets in 10 languages: English, Spanish, Italian, Portuguese, French, Chinese, Hindi, Arabic, Dutch and Korean, of which, the last 4 are not in the training data. Our approach to address this task is based on data augmentation and the use of three multilingual Large Language Models (multilingual BERT, XLM and mDeBERTA) by ensemble learning. Our team ranked 30th out of 45 participants. Our best results were achieved with two unseen languages: Korean (16th) and Hindi (19th).

1 Introduction

In Natural Language Processing (NLP), intimacy can be described as how people communicate their perception and willingness to share personal data and emotions to their audience (Pei and Jurgens, 2020). The SemEval 2023 Task 9, entitled Multilingual Tweet Intimacy Analysis (MTIA) (Pei et al., 2023), consists of a regression task in which the participants should rate in a score from 1 to 5 the intimacy of short documents written in 10 languages: English, Spanish, Italian, Portuguese, French, Chinese, Hindi, Arabic, Dutch and Korean. This task was co-organized by University of Michigan and Snap Inc. There are two main challenges concerning this task. On the one hand, the training dataset provided to the participants does not cover all the evaluated languages, but only six of them: English, Spanish, Italian, Portuguese, French, and Chinese. However, the evaluation is conducted in those six languages plus Hindi, Arabic, Dutch and Korean. On the other hand, participants were only allowed to submit a unique run, which hinders the shared task.

Our strategy to solve the MTIA challenge consists of an ensemble learning composed of three multilingual Large Language Models (LLM): multilingual BERT (Devlin et al., 2018), XLM (Lample and Conneau, 2019), and mDeBERTA (He et al., 2021). Besides, we use data augmentation incorporating to the training the dataset suggested by the organizers and provided in the work of Pei and Jurgens (2020), with more than two thousand English questions from Reddit and other sources and annotated with intimacy scores in the range [-1, 1].

Our participation achieved modest results in the task, reaching the 30th position in the leader-board, with a Pearson's R of 0.53. The best result is achieved by Lazybob, with a Pearson's R of 0.62. As commented above, as the participants were only allowed to submit a unique run, the analysis of our proposal is mainly based on a custom validation split. Additional resources concerning our participation can be found at `https://github.com/NLP-UMUTeam/semeval-2023-mtia`.

2 Background

The organisers of the task provided the participants with the novel MINT dataset (Pei et al., 2023), whose original training split consists of 9491 tweets rated with an intimacy score. The tweets were compiled between 2018 and 2022. To obtain tweets in different languages, the authors combined language filters in Twitter with language detectors models such as fastText (Joulin et al., 2016). Next, the authors created clusters of tweets of each language and several annotators rated the tweets in a scale from 1 (not intimate at all) to 5 (very intimate). As it can be observed, in the histogram plotted in Figure 1, most of the samples are rated with low scores. Regarding the six languages involved during the training, these are almost balanced, with 1596 documents written in Portuguese and Chinese,

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 293–299
July 13-14, 2023 ©2023 Association for Computational Linguistics

1592 in Spanish, 1588 in French, 1587 in English and 1532 in Italian. An example of the dataset is the Spanish text "Necesito paz mental"[1], rated with an intimacy score of 2.8. In Figure 2 the rounding label distribution is shown. The majority of labels are between 2 and 3 and with fewer instances of labels near to 0 or 5.

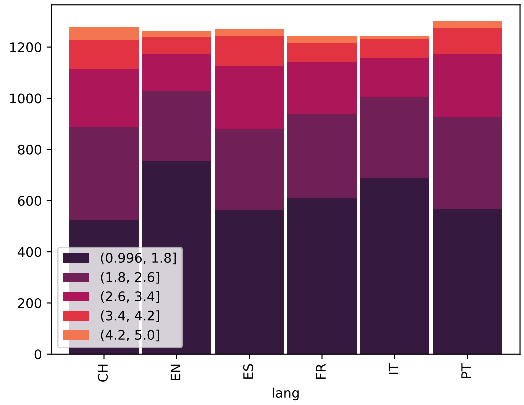

Figure 1: Histogram of the Intimacy score over the dataset, grouped per language

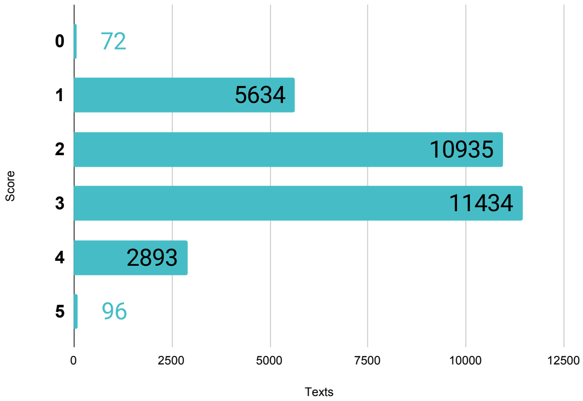

Figure 2: Rounding label distribution

The participants of the task were encouraged to use the dataset provided in Pei and Jurgens (2020); which contains English sentences with an intimacy score between -1 and 1.

3 System Overview

Our pipeline for solving the MTIA 2023 shared task is depicted in Figure 3. In a nutshell, it can be described as follows. First, we clean and pre-process the MTIA dataset and keep a small portion of the training split to create a custom validation split. Second, we perform a data augmentation stage applying Google Translate to the dataset of Pei and Jurgens (2020). Third, we evaluate three multi-lingual LLMs and one model based on linguistic features. Forth, we build an ensemble learning model that averages the predictions of the three LLMs to send our final predictions to the organizers of the task.

Concerning the data cleaning stage, we strip hyperlinks, hashtags, mentions and white space characters. Regarding the dataset splitter step, we reserve a 20% of the tweets from the training split for custom validation purposes. Next, we enlarge the training dataset proposed by incorporating the dataset provided in Pei and Jurgens (2020). This dataset contains sentences written in English. We use Google Translate to translate these sentences to Spanish, Italian, Portuguese, French, Hindi, Arabic, Dutch and Korean. This way, we could incorporate 21573 new sentences to the training. As this dataset is rated in rank from -1 to 1, we translate the ratings to a scale from 1 to 5, maintaining the ratio. Besides, it is worth noting that none of these new instances are used for custom validation.

4 Experimental Setup

During the evaluation phase, apart from the multilingual LLMs, we evaluate the usage of linguistic features from UMUTextStats (García-Díaz et al., 2022c). The linguistic features from UMUTextStats have been evaluated in several NLP tasks, such as author profiling (García-Díaz et al., 2022a), satire identification (García-Díaz and Valencia-García, 2022), and hate-speech detection (García-Díaz et al., 2022b).

UMUTextStats is designed for the Spanish language, but it has a subset of language-independent features. These features are stylometric features and features related to Named Entity Recognition (NER) and Part-of-Speech (PoS). To extract these features, UMUTextStats relies on Stanza (Qi et al., 2020) to extract some features related to PoS and NER. However, not all the languages involved in the MTIA shared task have models per Stanza, so the linguistic features were not useful for some of the languages involved on this shared task. Accordingly, we decided not to include the Linguistic Features (LF) in the final submission.

However, we use the linguistic features to make an analysis of the Spanish split of the dataset and we observe a correlation with misspelled words

[1] In English: I need peace of mind

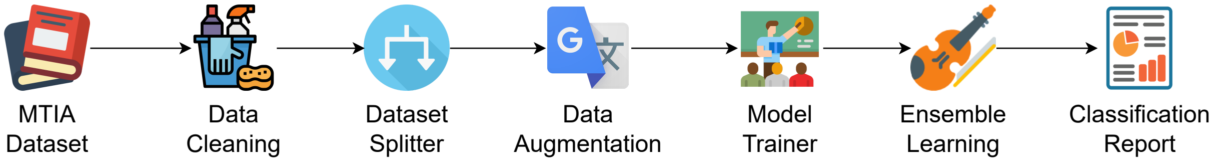

Figure 3: System architecture

with intimacy followed by morphological features related to proper and common nouns, personal pronouns in first, second person, and third person. We also identify a correlation with stylometric clues concerning the length of the tweets and with the usage of hyperboles, proper from figurative language. These results are depicted in Figure 4.

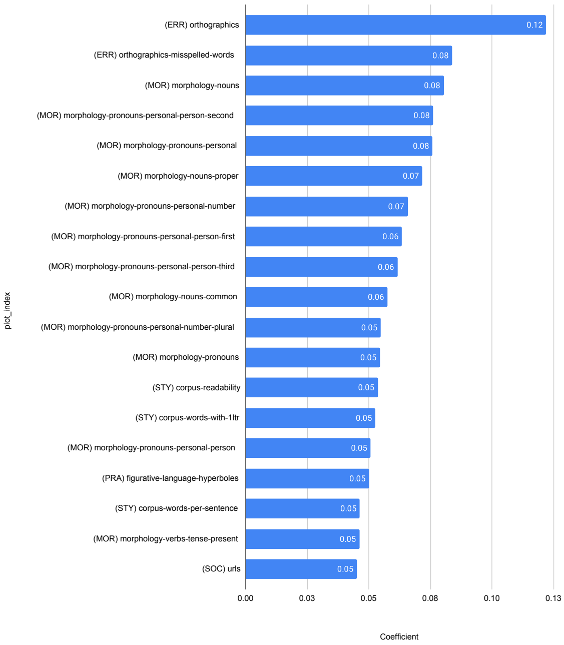

Figure 4: Information gain of the best 20 linguistic features

Next, the regression neural network architecture is described. For each LLM we conduct an hyperparameter optimization stage consisting of training 10 models for each LLM evaluating different parameters, including the learning rate, the number of epochs for traning, the warm up steps and the weight decay. The results of the best model for each LLM are depicted in Table 1. It can be observed that the best models require for small training epochs and all require of warmup steps and weight decay. Next, we extract the sentence embeddings for each LLM. This results in a vector of size 768 for each document.

Finally, we conduct another hyperparameter optimization stage using Keras. We follow this step to be consistent with the LFs and the LLMs. The results of this experiment are reported in Table 2. For each feature set, we evaluate 55 models changing the neural network architecture (its number of neurons and hidden layers), the dropout, the batch size and the activation function. We can observe that the best models for the LF and for mBERT are complex neural networks with 5 and 8 hidden layers respectively. The LF neural network has brick size (all layers have the same number of neural networks) but mBERT has a diamond shape (the inner layers have much more neurons). All models benefit for a strong dropout mechanism and most of them also benefit from large batch sizes.

5 Results

5.1 Validation results

The goal of the MTIA shared task is to predict the intimacy of tweets with a range from 1 to 5. The performance of each submission is ranked based on Pearson's R over the test split. However, for our analysis with our custom validate split, we evaluate the following metrics: (1) Explained Variance (EV), that measures how many information we lose by approximating the dataset. A small EV indicates the training process has strong oscillations; (2) Root Mean Squared Logarithmic Error (RMSLE), is the root mean squared error of the log-transformed predicted and log-transformed real values. RMSLE is an effective metric when the label has exponential growth and when we want to measure the percentage of errors, instead of the absolute value of errors; (3) Pearson's R, measures the strength of the linear association between the predicted and the real values. Pearson's R is the official metric for the MTIA shared task; (4) R-squared (R2), which is the proportion of variation in the outcome that is explained by the predictor variables. The higher the R-squared, the better the model; (5) Mean Absolute Error (MAE), that is the average absolute difference between the predicted and real values. MAE is less sensitive to outliers; (6) Mean Squared Error

Table 1: Hyperparameter optimization stage of the LLMs

LLM	Learning Rate	Training Epochs	Warmup steps	Weight decay
MBERT	2e-05	1	500	0.034
XLM	3.3e-05	2	500	0.28
MDEBERTA	2.7e-05	1	250	0.081

Table 2: Hyperparameter optimization stage of the feature sets

	Shape	Layers	Neurons	Dropout	Batch size	Activation
LF	brick	5	8	0.2	64	elu
MBERT	diamond	8	37	0.2	64	tanh
XLM	brick	2	37	0.3	64	relu
MDEBERTA	brick	2	128	0.3	32	relu

(MSE), which is the average squared difference between the observed real values and the predictions of the model; and the (7) Root Mean Squared Error (RMSE), which is the square root of MSE. The lower the RMSE and MSE, the better the model.

Table 3 contains the results with the custom validation split. The ensemble model of MBERT, XLM and MDEBERTA is the best model concerning all the evaluated metrics, reaching 0.46 of EV, 0.042 of RMSLE, 0.46 of R2, 0.516 of MAE, 0.426 of MSE and 0.652 of RMSE. As expected, the model based on linguistic features (LF) is the most limited model, as not all the evaluated languages contain models in Stanza to get the set of language independent variables. Accordingly, we decided not to include the LF in the final ensemble, as it would decrease the overall performance of the ensemble. Out of the LLMs evaluated, a relevant different is found between multilingual BERT and the other two multilingual LLMs. However, as the ensemble learning method improves the results of the three LLMs, we consider that these LLMs complement each other.

5.2 Official leader board

The test split of MTIA 2023 consists of 13697 sentences. Similar to the train split, the testing split is also almost balanced, from Hindi, with 1260 tweets to Korean, with 1410 tweets.

A total of 45 teams participated in the MTIA 2023 shared task. Table 4 contains the official leader-board. For the sake of simplicity, we only include here the top 5 teams, our result and the result of the last position. The average Pearson's R of all participants is 0.5105797444 with a standard deviation of 0.145161199. As it can be observed, our proposal based on ensemble learning scored 0.532, which is slightly superior of the average but far from the top-five scores.

Our result in the official leader board (0.532 of Pearson's R) is more limited than our result with our custom validation split (0.682 of Pearson's R). It is possible that this difference is due to the unseen languages that are incorporated for the test split. The results of our proposal by language are depicted in Table 5. These results are organized into two groups. The first group are the six languages used during the training and the second group are the 4 unseen languages during the training. It can be observed that our best result is achieved with an unseen language, Korean, reaching position 16. Our second best score is also with another unseen language, Hindi. However, we got limited results for Portuguese, Dutch, and Arabic.

5.3 Error Analysis

In order to understand visually the classifications, we rounded the predictions to create a confusion matrix (see Figure 5). As we can observe, the model has good performance with the ratings that are in the rank between 1 and 3, being the ratings equal or higher than 4 the ones with major limitations. It seems that our ensemble model assigns higher values than the actual ones. For example, 347 instances with scores near to 1 were assigned to the bin of labels near to 2, and 141 labels with scores near to 2 were assigned to the bin of labels near to 3. However, out of the 144 tweets with scores near to 4, the ensemble model assigns scores near of 3 to 101 tweets, and scores near to 2 to 33

Table 3: Results with the custom evaluation split. Reporting the Explained Variance (EV), Root Mean Squared Logarithmic Error (RMSLE), Pearson's R, R-squared (R2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)

	EV	RMSLE	PEARSON'S R	R2	MAE	MSE	RMSE
LF	0.197	0.062	0.444	0.197	0.640	0.633	0.796
MBERT	0.374	0.048	0.612	0.374	0.555	0.494	0.703
XLM	0.432	0.044	0.658	0.432	0.524	0.448	0.669
MDEBERTA	0.449	0.043	0.670	0.449	0.516	0.434	0.659
ENSEMBLE	**0.460**	**0.042**	**0.682**	**0.460**	**0.516**	**0.426**	**0.652**

Table 4: Top 5 results of the leaderboard compared with our result and the result of the last position

Team	Score	Ranking
lazybob	0.616	1
UZH_CLyp	0.614	2
opi	0.613	3
tmn	0.599	4
OPD	0.599	5
UMUTeam-SINAI	**0.532**	30
uaic_mt_2023	0.004	45

Table 5: Detailed results per language

Language	Score	Ranking
English	0.642	33
Spanish	0.705	26
Portuguese	0.582	35
Italian	0.659	31
French	0.611	35
Chinese	0.704	23
Total	0.664	31
Hindi	0.220	19
Dutch	0.539	35
Korean	0.362	16
Arabic	0.503	32
Total	0.360	32
Total	0.532	30

tweets, leaving only 10 tweets correctly classified. These results suggest that our model is not suitable for tweets with larger intimacy scores.

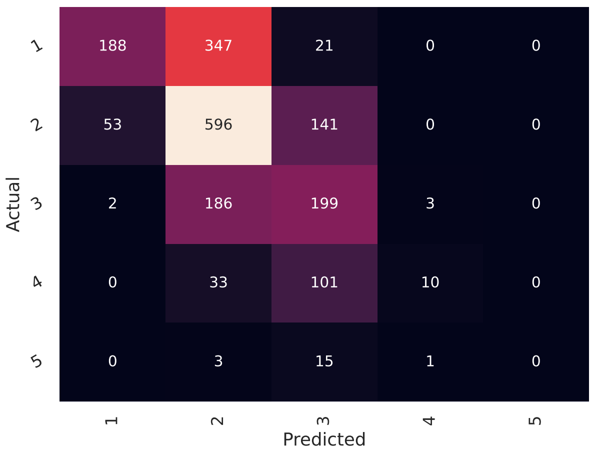

Figure 5: Confusion matrix with the validation split with the ensemble model

5.4 Ablation Analysis

To understand the contribution of the data augmentation stage in our pipeline, a experiment with the custom validation split but without the augmented data is performed. The results are reported in Table 6 in which the difference between both experiments are shown. As some of the metrics are *the lower, the better* we have included the symbols ↑, ↓, and - to indicate when the effect of data augmentation improves, downgrades or it does not have effect in the performance. As it can be observed, the data augmentation step of our pipeline is beneficial for the performance, but not for all the experiments. However, the contribution is not very high. For example, there is only a difference of 0.0009 with the RMSE for the ensemble, between the experiments with and without data augmentation.

Table 6: Ablation analysis of the data augmentation with the custom evaluation split. We report the difference between the results achieved with and without data augmentation. The ↑ symbol indicates that data augmentation improves the performance whereas ↓ indicates the performance decreases. The "-" symbol denotes no effect at all. The metrics are the Explained Variance (EV), Root Mean Squared Logarithmic Error (RMSLE), Pearson's R, R-squared (R2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

	EV	RMSLE	PEARSON'S R	R2	MAE	MSE	RMSE
LF	0.003 ↑	0 -	0.002 ↑	0.004 ↑	0.008 ↓	-0.003 ↑	-0.002 ↑
MBERT	-0.004 ↓	0.001 ↓	-0.003 ↓	-0.004 ↓	0.012 ↓	0.003 ↓	0.003 ↓
XLM	0.001 ↑	0 ↓	0.001 ↑	0.001 ↑	-0.002 ↑	0 -	-0.001 ↑
MDEBERTA	- 0.006 ↓	0.001 -	-0.005 ↓	-0.006 ↓	0.004 ↓	0.004 ↓	0.003 ↓
ENSEMBLE	0.013 ↑	0 -	-0.004 ↓	0.014 ↑	-0.005 ↑	-0.01 ↑	-0.009 ↑

6 Conclusion

Despite the fact that our results are limited, we are very pleased with our participation. First, because this is the first time we participated in a shared task concerning intimacy. Second, because the MTIA shared task was challenging as we could only send one result and because there are four unseen languages during testing. Our proposal based on ensemble learning on three multilingual LLM reached position 30th in the official leaderboard from a total of 45 participants. Our best results are achieved with two unseen languages: Korean (16th) and Hindi(19th).

After the evaluation of our results, we consider that there are several ways in which we could have improved our results. First, we should have conducted an in-deep analysis of the dataset. However, this was not easy for us because we are not fluent speakers of many of these languages, so we can miss important aspects related to the context. Second, it is possible that the data augmentation process was not beneficial for the performance of our model, as the translations could be less accurate in some languages or it is possible that cultural and background differences are not well represented in the dataset. However, we consider that we could have translated all sentences into a common language (Spanish or English, for instance) and could include features related to topics to our model. We will explore this path in future multilingual shared tasks. Three, our models could be biased to our custom validation split. In this sense, we will incorporate to our pipeline a nested-cross validation evaluation. Fourth, our ablation analysis is limited, as we only consider the data augmentation step. However, we need to conduct more experiments in order to gain understanding of other modules such as the preprocessing module.

Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme, and Salud María Jiménez-Zafra is partially supported by a grant from Fondo Social Europeo and Administración de la Junta de Andalucía (DOC_01073).

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2022a. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022b. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

José Antonio García-Díaz, Pedro José Vivancos-Vicente, Ángela Almela, and Rafael Valencia-García. 2022c. UMUTextStats: A linguistic feature extraction tool for Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6035–6044, Marseille, France. European Language Resources Association.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326. Association for Computational Linguistics.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. Semeval 2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D Manning, Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2020. Stanza: A {Python} natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.