BERTbek: A Pretrained Language Model for Uzbek

Elmurod Kuriyozov[1,2], **David Vilares**[1] **and Carlos Gómez-Rodríguez**[1]
[1]Universidade da Coruña, CITIC, Grupo LYS, Depto. de Computación y Tecnologías de la Información,
Facultade de Informática, Campus de Elviña, A Coruña 15071, Spain
[2]Urgench State University, Department of Computer Science,
14, Khamid Alimdjan street, Urgench city, 220100, Uzbekistan
{e.kuriyozov, david.vilares, carlos.gomez}@udc.es

Abstract

Recent advances in neural networks based language representation made it possible for pretrained language models to outperform previous models in many downstream natural language processing (NLP) tasks. These pretrained language models have also shown that if large enough, they exhibit good few-shot abilities, which is especially beneficial for low-resource scenarios. In this respect, although there are some large-scale multilingual pretrained language models available, language-specific pretrained models have demonstrated to be more accurate for monolingual evaluation setups. In this work, we present BERTbek - pretrained language models based on the BERT (Bidirectional Encoder Representations from Transformers) architecture for the low-resource Uzbek language. We also provide a comprehensive evaluation of the models on a number of NLP tasks: sentiment analysis, multi-label topic classification, and named entity recognition, comparing the models with various machine learning methods as well as multilingual BERT (mBERT). Experimental results indicate that our models outperform mBERT and other task-specific baseline models in all three tasks. Additionally, we also show the impact of training data size and quality on the downstream performance of BERT models, by training three different models with different text sources and corpus sizes.

Keywords: BERT, language modeling, Uzbek language, natural language processing; low-resource languages

1. Introduction

The approaches towards natural language processing (NLP) applications have seen a rise in pretrained large language models (LMs) on large unlabeled data to solve downstream NLP tasks over the last years. These pretrained LMs are then usually used in zero-shot or few-shot setups, being fine-tuned to fit the LM output to a specific NLP task, often achieving state-of-the-art performances (Radford et al., 2018; Devlin et al., 2019; Yang et al., 2019; Lample and Conneau, 2019). One of the most popular approaches used to create these LMs relies on using Transformers-based architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019), GPT (Radford et al., 2018), as well as XLM (Lample and Conneau, 2019), among many others. Especially, BERT has been particularly influential, due to its early adoption and success in a range of downstream NLP tasks in English and other languages.

Along with monolingual models, multilingual models have been developed for the same kind of architectures, like multilingual BERT, XLM (Lample and Conneau, 2019), and XLM-RoBERTa (Conneau et al., 2019). These multilingual models are interesting because they have been proven to perform well for cross-lingual transfer-learning (Wu and Dredze, 2019). However, they also have some problems: (1) Multilingual pretrained LMs could not outperform their monolingual counterparts in monolingual evaluation settings (Virtanen et al., 2019; Safaya et al., 2020; de Lima et al., 2022); (2) Multilingual language models require larger vocabulary size and number of training parameters, thus requiring more GPU performance and time to fine-tune them; (3) Creating LMs trained on quality data is important for reliable evaluation (Melis et al., 2017; Xu et al., 2022), especially when the size and diversity of non-English data involved are considered in pretraining multilingual models (Pires et al., 2019).

Apart from the fact that these neural pretrained LMs are favored in terms of their better performance, they can be pretrained just on raw texts, reducing the reliance on large amounts of labeled data, which works in favor of low-resource scenarios where such data is scarce (Kryeziu and Shehu, 2022). For the above-mentioned reasons, besides English, monolingual BERT models have been trained for different languages: rich-resourced ones such as Spanish (Canete et al., 2020), Russian (Kuratov and Arkhipov, 2019), and Portuguese (Souza et al., 2020); as well as low-resource languages like Galician (Vilares et al., 2021), Maltese (Micallef et al., 2022), Armenian, Kazakh, or Tamil (Tsai et al., 2019).

In this work, we present BERTbek - openly available pretrained BERT-based language models for Uzbek, a low-resource language like the majority of other counterparts in the Turkic family. We

SIGUL2024 Workshop, pages 33–44

first collect raw text corpora from different sources like Wikipedia and news websites, then pretrain BERT language models with different text sources and sizes. We also evaluate the models performance in number of downstream NLP tasks, such as sentiment analysis, multi-label text classification, and named entity recognition, against various task-specific baseline models, including multilingual BERT. Our experiments indicate that not only the size, but also the quality and source of the training text directly affect the downstream performance of the pretrained models. Also, BERTbek monolingual models not only outperform their multilingual counterpart, but also other task-specific neural models without pretraining in all the evaluated tasks. All the code used in this work is openly available at the project's GitHub repository[1] and the BERTbek models have been uploaded to the HuggingFace Models Hub[2].

2. Related Work

The evolution of current transfer learning techniques dates back to word (or sub-word) level vector representations, such as word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017), among the most popular models for generating static word embeddings. These models were trained on large unlabeled language corpora using shallow neural networks (Bengio et al., 2000; Collobert and Weston, 2008). A limitation of these traditional techniques is that they could only encode non-contextualized word representations, which is an issue to describe words with same spellings (homographs), words that have different meanings based on the context they appear in (polysemous), or simply to model rich in-context representations for words within a sentence. This was addressed by the more advanced methods proposed, for instance, by ELMo (Peters et al., 2018) and Flair (Akbik et al., 2018) embeddings, which use recurrent neural network (RNN) architectures to obtain richer context-sensitive embeddings.

More recently, word vector contextualization has shifted towards large pretrained LMs with deep transfer learning techniques, after the successful introduction of the attention-based Transformer (Vaswani et al., 2017) architecture. One popular example is the BERT model presented by Devlin et. al (Devlin et al., 2019), a bidirectional encoder representation model using Transformers. For pretraining, BERT models optimize two language objectives, namely masked language modeling (MLM) and next sentence prediction (NSP), where the former training objective tries to predict a word hidden with a special label ([MASK]) in a given sentence (also known as Cloze task), and the latter predicts the logical or contextual connection between two sentences.

The success of the BERT model that was originally trained in English together with its multilingual variant (mBERT, trained using more than a hundred languages in one big model) has also attracted attention from research communities in other languages. As a result, a number of monolingual pretrained BERT models for many other languages were released, e.g., Russian (Kuratov and Arkhipov, 2019), Arabic (Antoun et al., 2020), Czech (Sido et al., 2021), or models for specific subdomains of English, such as medical sciences (Lee et al., 2020), or finance (Yang et al., 2020), to name a few. Also, various studies have taken place to study the way in which BERT-based models encode the language knowledge in its deep architecture (Lin et al., 2019; Ettinger, 2020), or syntax-sensitive phenomena (Vilares et al., 2020).

Furthermore, a number of successors of BERT were proposed with various optimization methods to the original model, while maintaining similar performance results. For instance, RoBERTa (Liu et al., 2019) proposes an improved recipe for training BERT models that suggests training on longer sequences and dynamically changing the masking pattern. The paper also reports that training the model with bigger data and for longer time improves the model performance on NLP benchmarks. Another recent work, called ALBERT (Lan et al., 2020), proposed a BERT-based model with lesser computational cost, by reducing the number of training parameters (25M less than the base model) that helps to both use less memory space and train faster. Performance enhancement was also achieved by introducing cross-layer parameter sharing and replacing the NSP training task with a sentence order prediction (SOP) one.

Regarding the focus language of this work, the Uzbek language is included in multiple multilingual pretrained LMs, such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2019), and mT5 (Xue and et, 2020), where the texts were collected from Wikipedia and CommonCrawl. Mansurov and Mansurov (2021b) developed a monolingual pretrained LM based on BERT architecture, named UzBERT, with very much like parameters as the original BERT-base model (12-layers, 110M parameters, 30K vocabulary size, MLM and NSP training objectives). UzBERT was pretrained using news corpus collected from websites in Uzbek language, covering various domains like economics, law, literature and agriculture, totalling around 140M words. A main downside of the UzBERT model is the choice of alphabet to

[1] https://github.com/elmurod1202/BERTbek

[2] https://huggingface.co/elmurod1202/bertbek-news-big-cased

collect training text, where authors used Cyrillic, which is an old alphabet of Uzbekistan with many websites, books, and even official documents still available (Salaev et al., 2023; Madatov et al., 2022). This leaves alternative space to create BERT-based language model for Uzbek, in particular in the official Latin script.

3. BERTbek Models

This section includes brief information about the Uzbek language and the steps taken to train the BERT models for Uzbek, like data collection, vocabulary creation and pretraining.

3.1. Uzbek Language

Uzbek (native: O'zbek tili) belongs to the Eastern Turkic or Karluk branch of the Turkic language family, also referred as Northern Uzbek language to not to mistake it with the Southern Uzbek, which is another variety of Uzbek spoken by an ethnic Uzbek minority in Afghanistan (which together with northern Uzbek, they form one macrolanguage). It is the only national and the first official language of Uzbekistan (Sharipov et al., 2022; Madatov et al., 2023). Uzbek is spoken by more than 30 million speakers inside Uzbekistan alone, and more than ten million elsewhere in neighbouring Central Asian countries, the Southern Russian Federation, as well as the North-Eastern part of China (Salaev et al., 2022b). Although it is the second most widely spoken language among Turkic languages (right after the Turkish language), it is considered as a low-resource language due to scarce availability of NLP resources and tools (Matlatipov et al., 2022; Sharipov and Yuldashov, 2022).

3.2. Training Data Collection

To provide a sufficiently large and varied text corpus for pretraining the BERT model, we collected Uzbek texts from two primary sources: Wikipedia and news data.

Wikipedia corpus. The Wikipedia corpus was collected from the Uzbek version of Wikipedia [3], more specifically, from the 2022-01-20 dump [4] with around 124K articles. For extracting raw text and cleaning, the `wikiextractor` tool [5] was used. Post-cleaning process was used to clean the collected texts as some of the articles in Uzbek Wikipedia contained words in Latin script with some of letters mixed with their homoglyphs [6] in Cyrillic. For this, we identified articles that contain homoglyphs in Cyrillic, and replaced with their correct alternatives in Latin. Although encyclopedic data, such as Wikipedia, are a common choice to create text corpus in NLP (Nothman et al., 2013; Virtanen et al., 2019; Vilares et al., 2021) for its coverage of various topics and genres, Uzbek Wikipedia has many articles that were created by bots that used either automatically translated text or articles generated from predefined structures. Another downside of this source is the fact that the majority of the Uzbek Wikipedia articles were bulk imported from Uzbek Encyclopedia (Aminov et al., 2000-2006) directly, which were written in a terse style with an abundance of abbreviations to save printing space (Mansurov and Mansurov, 2021b). All these factors mentioned above result a corpus with a lower data quality.

News corpus. The News corpus was collected from 'Daryo' [7], the most popular news portal in Uzbekistan [8], using the Scrapy web crawler tool [9]. Around 200K articles were collected from Daryo news in various domains, such as sport, tech, law, economics, health, etc. Daryo offers the same news article in two scripts, Cyrillic and Latin, we collected only Latin ones. For only the minority of the news data that were not available in the Latin alphabet, we collected the Cyrillic ones, and transliterated them into the Latin scheme using a Python machine transliteration tool for Uzbek (Salaev et al., 2022a). This collection of texts serves as a good quality corpus, due to the structural variety and complexity of the sentences, and the cleanliness of the texts contained in it compared to the Wikipedia corpus. We also decided to use this news data in two forms, first we took all of the collected data (around 200K articles) and named it as 'News-big', then we took another smaller part of it (around 56K articles) that was cut down to the size of our Wikipedia corpus (both having roughly 9.7M tokens) and named it 'News-small'. Overall, having these Wikipedia and two forms of news corpora allows us to use them for training three different BERT models and achieving this work's two main goals: (i) Compare how data quality affects over models trained with two corpora of the same size (using Wikipedia and News-small); (2) Analyse how the training data size

[3] https://uz.wikipedia.org/wiki

[4] https://dumps.wikimedia.org/uzwiki

[5] https://github.com/attardi/wikiextractor

[6] *Homoglyph* (a term from ortography or typograpy) is one of two or more characters, with shapes that appear identical or very similar. In the case of Uzbek Wikipedia, it was caused by bad transliteration practice from Cyrillic to Latin when creating articles.

[7] https://daryo.uz

[8] https://www.uz/uz/stat/visitors/ratings

[9] https://scrapy.org

affects the model performance over two models trained on the same data source but different sizes (using News-big and News-small).

In both corpus sources, the titles were also included alongside the article body. To make sure that none of the texts used for evaluation were not seen during the training the BERT models, all the sentences used in the sentiment analysis and named entity recognition experiments (these experiments are explained thoroughly in Section 4) were removed from all three corpora. More about the detail size comparisons of all corpora can be seen in Table 1.

Table 1: Number of articles, sentences and tokens in each corpus.

Corpus name	Articles	Sentences	Tokens
Wikipedia	120K	2M	9.7M
News-small	56K	0.8M	9.7M
News-big	190K	2.6M	32.5M

3.3. Pretraining

Here we explain the steps taken for vocabulary generation and pretraining the BERT models.

3.3.1. Vocabulary Generation

Pretraining a language model requires a vocabulary of sub-word pieces with a set size for a language to tokenize training texts using that vocabulary, where most common tokens are described in one piece, lesser common ones can be described using a combination of smaller word-pieces, and the least common or not seen ones get a specified label (UNK). We generated a dedicated BERT vocabulary for Uzbek, by gathering all raw data we collected (Wikipedia and news) and tokenized it using BERT WordPiece tokenizer, following the same setup that was used in the original English tokenizer. We use cased vocabulary, since casing is an important aspect for some NLP tasks, such as the named entity recognition task we use in the experiments. For the size of the vocabulary, we chose 30K word pieces, following the common practice of other monolingual BERT models, like English (Devlin et al., 2019), Spanish (Canete et al., 2020), or Russian (Kuratov and Arkhipov, 2019). Similar vocabulary size (32K) was also used by Turkish BERT [10], a language in the same family. For this reason, we use the vocabulary with the same size (30K) further in all training and experiments in this work, leaving the topic of finding the optimal vocabulary size and its effect on the model performance for Uzbek and other Turkic languages for a future work. We set the minimum frequency limit of the vocabulary down to two, because of the agglutinative nature of Uzbek where words are used in various inflectional and derivational forms, hence lowering the word-form frequency.

[10] https://github.com/stefan-it/turkish-bert

3.3.2. Pretraining Parameters

As determining the impact of training data size and quality to the overall BERT model performance is one of the key contributions of this work, we trained three different BERTbek models with different data sources and sizes, which are named as follows:

- $BERTbek_{Wiki}$ model, trained using around 120K articles extracted from Uzbek Wikipedia;
- $BERTbek_{News-Small}$ model, trained using news corpus, limited to only 56K articles (containing the same number of tokens as the previous $BERTbek_{Wiki}$ one);
- $BERTbek_{News-Big}$ model, trained using the same news corpus, but with all 190K articles collected from Daryo.

The first 95% of the texts were taken as a training set and the remaining 5% were used as a dev set in all three cases. In the case of news corpus, the domains of the articles (new categories) were also considered to provide the same diversity for both sets. For most of the training hyperparameter setup, and all of the codes used, we followed the original BERT paper for all three models. We trained models on Masked Language Modeling (MLM) task using 12 transformer layers, 768 hidden dimensions and 12 attention heads. 30K size of vocabulary described above was used for the tokenizer. The Adam optimizer with decoupled weight decay (Loshchilov and Hutter, 2017) was used with a learning rate set to 1e-4 with 10,000 warm-up steps.

The *transformers* library by HuggingFace (Wolf et al., 2020) was used to train each model using a PC with two NIVIDIA GeForce RTX 3090 GPUs (24GB each) for around 18 days until they reached 3M steps (the $BERTbek_{News-Big}$ model was later trained further to assess the performance gain, this will be discussed in Section 5).

4. Experiments and Results

This section describes the evaluation results of the pretrained BERTbek models by fine-tuning them for three different downstream NLP tasks, namely sentiment analysis, topic classification, and named entity recognition. We fine-tuned the models pretrained in the previous section for our target tasks. For this step we again used specific classes provided by the *transformers* library (unless explicitly

stated, default parameters were used) and the training and dev sets of datasets were used for fine-tuning.

4.1. Datasets for Downstream Tasks

Sentiment analysis. The dataset we used for this evaluation task was obtained from the work of Kuriyozov et.al. (Kuriyozov et al., 2022), in which the authors present two datasets: the first comprises about 4.5K reviews extracted from Google's Android app store [11] reviews in Uzbek and manually annotated (hence called `"Manual dataset"`); and the second dataset is automatically translated from around 8.5K movie reviews in English into Uzbek, with minor manual corrections (and called `"Translated dataset"`). Both datasets are annotated with binary sentiment classification (positive and negative labels for each review).

The splits provided for both datasets in the original paper were only training and test sets, but no development one, so we redivided the datasets and split them into train, dev, and test splits with 0.5 x 0.2 x 0.3 ratio, respectively, to use the dev set for fine-tuning.

Topic classification. There is no officially available multi-label text classification dataset for the Uzbek language, so we followed the dataset creation methodology of Rabbimov et.al (Rabbimov and Kobilov, 2020) and created a new one from our news corpus. The Daryo news articles come with metadata that indicate what news category each article belongs to. There are more than 50 different categories associated with various amounts of articles in the corpus. We regrouped the articles by merging the smaller article categories in the same domain into one big category (like 'Auto', 'Gadgets', 'Technology' were grouped as one 'Tech' category, and 'Show-business','Cinema','Music' were grouped as 'Media', etc.), to simplify the dataset with labels down to ten, and also helping to reduce the imbalance between the samples of different categories.

Also, when choosing articles to create a dataset for this task, we made sure that no article appears as a source of BERTbek model pretraining in at least two models ($BERTbek_{Wiki}$ and $BERTbek_{News-small}$ models), which we used for evaluation. The detailed information regarding all the news categories, as well as the number of articles are reported in Table 2.

We split the created dataset into train, dev, and test sets with 0.5 x 0.2 x 0.3 ratio, respectively. We also made sure that each set would get news texts equally distributed over all the categories.

Named Entity recognition. For this task we use the UzNER dataset [12] that consists of 300 news articles with around 95K tokens in total, balanced over ten different domains, such as Sport, Tech, Media, Science, etc. The same news text source as our news corpus was used for the UzNER dataset and it contains roughly 7K named entities (12% of the overall tokens in the dataset) over six named entity labels: Organisation (ORG), person (PER), location (LOC), date (DATE), time (TIME), as well as miscellaneous (MISC). We use the original splits provided by the dataset with training, evaluation, as well as testing sets with 0.5 x 0.2 x 0.3 ratios, respectively.

Table 2: Names, number of articles, and names of subcategories included per category.

Category	Articles	Category	Articles
Local	49404	**Media**	3067
World	43909	**Culture**	3040
Sport	19375	**Science**	1541
Tech	8470	**Health**	889
Misc	3318	**Food**	405
TOTAL:		**133418**	

4.2. Baseline Models

We use mBERT (official base model [13]) as a baseline model to compare the performance results in all three tasks. The other models used for each specific task are described below.

Text classification tasks. We evaluate from traditional bag-of-words models to sequential bidirectional neural network architectures. We applied the same methodology to both datasets, only difference being the number of labels to be predicted for each one: for the sentiment analysis task, we used a dataset with two labels (positive and negative), whereas the topic classification dataset we generated from the news texts uses ten different labels.

More specifically, the baselines used for comparison are:

- $LR_{Word-ngrams}$: Logistic regression with word-level n-grams (unigram and bi-gram bag-of-words models, with TF-IDF scores);
- $LR_{Character-ngrams}$: Logistic regression with character-level n-grams (bag-of-words model with up to 4-character n-grams);
- $LR_{Word+Char-ngrams}$: Logistic regression with word and character-level n-grams (con-

[11] https://play.google.com/store/apps

[12] The UzNER dataset was taken from a work that is not publicly available yet. We will share this information upon acceptance in the Appendix.

[13] https://huggingface.co/bert-base-multilingual-cased

catenated word and character TF-IDF matrices);

- RNN: Recurrent neural network without pretrained word embeddings (bidirectional GRU with 100 hidden states, the output of the hidden layer is the concatenation of the average and max pooling of the hidden states);
- $RNN_{Word-embeddings}$: Recurrent neural networks with pretrained word embeddings (previous bidirectional GRU model with the SOTA 300-dimensional FastText word embeddings for Uzbek (Kuriyozov et al., 2020));
- CNN: Convolutional neural networks (multi-channel CNN with three parallel channels, kernel sizes of 2, 3 and 5; the output of the hidden layer is the concatenation of the max pooling of the three channels);
- $RNN + CNN$: RNN + CNN model (convolutional layer added on top of the GRU layer);

For the detailed description of methodology setups, parameters, and the code of the above-mentioned models, readers are advised to refer to the original sentiment analysis dataset paper (Kuriyozov et al., 2022). That paper also presents evaluation results for these baseline models, but we cannot compare those results with our models' performance, since we used different splits. For this reason, we reproduced all the methods and calculated results using the same splits we used for our model evaluations.

All three BERTbek models were used for evaluation in the sentiment analysis task, but we skipped out the $BERTbek_{News-big}$ model in the topic classification task to provide a fair comparison, since the dataset of the latter task was part of its pretraining text source.

Named entity recognition. Besides multilingual BERT (mBERT), we also compare the BERTbek models' performance using following models with neural network architectures, as baseline models for this task:

- $LSTM_{Word}$: Word sequence layer with bi-directional LSTM encoder;
- $LSTM_{Char+Word}$: Word sequence layer on top of charaqter sequence layer, using bi-LSTM for both layers;
- $LSTM_{Char+Word} + W.emb.$: Character and word bi-LSTM sequence layers (as previous) with external pretrained word embeddings;
- $LSTM_{Char+Word} + W.emb. + CRF$: Character and Word bi-LSTM sequence layers with pretrained word embeddings (as previous) and CRF output layer;

The ($LSTM_{Word}$) model uses a single layer, the rest of the baseline models use two neural sequence layers of bi-directional long short-term memory (LSTM) encoder. Since it is bi-directional, both the left-to-right and right-to-left sequence information are captured, and the final two hidden states are concatenated. Character sequence layer takes character embeddings as an input, while word sequence layer takes character sequence representations (output of the previous layer) concatenated with word embeddings. Word embeddings are randomly initialized in the case of the first two models ($LSTM_{Word}$ and $LSTM_{Char+Word}$), but starting from $LSTM_{Char+Word} + W.emb.$ model, they are replaced by pretrained Uzbek FastText word embeddings (Kuriyozov et al., 2020). The $LSTM_{Char+Word} + W.emb. + CRF$ model has the same setup as the previous one, only with CRF output layer instead of softmax. All the models were built, trained and evaluated using NCRF++ [14] neural sequence labeling toolkit. The rest of the model setup, such as embedding sizes (word_emb_dim=300, char_emb_dim=30), training parameters (Adam optimizer for all models but $LSTM_{Char+Word}+W.emb.+CRF$ one, which uses SGD) as well as hyperparameters (learning rates, hidden dimensions, dropouts) were chosen according to the best performance using an evaluation performed on the development set.

4.3. Results

Sentiment analysis. The results of the sentiment analysis experiment are reported in Table 3. All three of our BERTbek models performed well in this task, outperforming the results of all but one of the methods previously studied by Kuriyozov et. al. 2022, and our $BERTbek_{News-Big}$ model has achieved the state-of-the-art results in both manual and translated datasets with 92.25 and 87.05 F1-scores, respectively. It is also worth mentioning that the RNN model performed better than BERTbek models in terms of precision score, but was low on recall, the opposite also applies to some other baseline models ($LR_{Word+Char-ngrams}$, $RNN + CNN$).

Topic classification. The evaluation results of BERT models for this task for all categories [15] is given in Table 4. Performance results (F1-score) for each category gives better understanding of how models perform based on each text domain, and its relation with the various sizes of the training data per label.

[14] https://github.com/jiesutd/NCRFpp

[15] Since the label attached to each document in the dataset is also the category name of the news article that makes up that document, we use terms 'label' and 'category' interchangeably in this task.

Table 3: Sentiment analysis evaluation results on two datasets: Manually collected app reviews of small size, and movie reviews translated from English with bigger size. F1-score (F1), Precision (Prec) and Recall (Rec) metrics are reported. The best performing model results for each metric are highlighted.

Model Name	F1_Manual	F1_Trans-d
$LR_{Word-ngrams}$	88.82	84.89
$LR_{Char-ngrams}$	90.38	85.78
$LR_{Word+Char-ngrams}$	91.97	86.39
RNN	88.19	84.69
$RNN_{Word-embeddings}$	90.01	85.54
CNN	89.38	85.24
$RNN + CNN$	90.67	85.70
$mBERT$	91.31	85.48
$BERTbek_{Wiki}$	91.14	85.74
$BERTbek_{News-Small}$	91.41	85.59
$BERTbek_{News-Big}$	**92.25**	**87.05**

The $BERTbek_{Wiki}$ model performs mostly on par with $mBERT$ due to the same source and similar size of Uzbek texts used for training, and the $BERTbek_{News-Small}$ model outperforms both in majority of the categories. Scores have a large variability range per category and all three models followed a similar pattern. The number of articles reported as reference indicates that not only the big size of documents enhances the performance results (the cases of 'Local' and 'World'), but also the uniqueness of the terminology used in the category context regardless of the limited availability of training data (like in the cases of 'Food' and 'Sport'). Moreover, the models struggled to predict the correct label for categories with wider domains that include various text contexts, in the cases of 'Misc', 'Media', and 'Science' categories.

Table 5 presents the overall evaluation results for topic classification, compared with the baseline models. The $BERTbek_{News-Small}$ model achieves the highest result in this task with a F1-score of 73.31, outperforming the next highest model result by at least 0.5 points ($RNN + CNN$). In terms of F1-score, although our $BERTbek_{Wiki}$ model (71.41) performed better than linear regression and mBERT models, it still lacked beind a couple of other baseline models, such as $RNN_{Word-embeddings}$ and $RNN + CNN$.

Named entity recognition. For all the evaluations in this task, we do not consider the non-entity tokens (labeled as "O"). The results indicate that the $BERTbek_{Wiki}$ model handled location (LOC) and time (TIME) entities better, while the $BERTbek_{News-Big}$ model performed best for organisation (ORG), person (PER), as well as miscellaneous (MISC) entities with F1-scores of 67.1, 91.2 and 58.57, respectively. Overall, all models achieve high scores for most of the entities, and the cases where models struggled can be explained by the very limited amount of entities appearing in the dataset (in the case of TIME, with only 45 entities in total), and the broad range of domains covered by a single entity (in the case of MISC, which includes all data regarding nationality, currency, percentage, metrics, etc.).

Overall NER results of all BERTbek and baseline models are reported in Table 6. In this task, only the $BERTbek_{Wiki}$ model achieved at least one point less score (for all metrics reported) than mBERT among all the tested BERT models. On the other hand, the $BERTbek_{News-Big}$ model has achieved the state-of-the-art results in this task with 78.69 F1-score, outperforming the next best model by at least 1.5 points.

5. Discussion

In this section, we discuss some of the tendencies the BERTbek models possess that were found in the evaluation tasks, such as the effect of pretraining data size and quality to the overall performance of BERT models.

Data size and quality. We trained two BERTbek models with the same training data size ($BERTbek_{Wiki}$ and $BERTbek_{News-Small}$ models) but different sources of text (Wikipedia and news data, see Section 3.2) to then analyse the models' performance. Although both models were trained using the same setups, the $BERTbek_{News-Small}$ model reached better results than the $BERTbek_{Wiki}$ one in all three NLP tasks we evaluated. Especially, it outperformed the alternative by at least two F1-score points in topic classification and NER tasks. This can be explained by a number of factors that lower the data quality of the Wikipedia corpus, such as many articles with the same structure that were created using bots as well as bulk import of articles from Uzbek Encyclopedia without correcting their terse style (Mansurov and Mansurov, 2021b). Overall, it can be inferred that data quality plays an important role in training BERT models.

Moreover, to analyse the performance differences of BERTbek models regarding training data size, two models were trained using the same text source and setups, but with different sizes: $BERTbek_{News-Small}$ and $BERTbek_{News-Big}$ models with around 10M and 32.5M tokens, respectively (reported in Table 1). As a result, the $BERTbek_{News-Big}$ model, that was trained using a corpus more than three times larger, outperformed not only other BERTbek models, but also all the other task-specific baseline models in all tasks we evaluated in this work, becoming the state-of-the-art model. This indicates that training

Table 4: Topic classification F1-scores for each news category for two of our BERTbek and multilingual BERT (mBERT) models. Number of articles per category is also reported for reference. Best scores per category are highlighted.

Models	Local	Tech	Misc	Sport	World	Media	Food	Health	Culture	Science
# of articles	49404	8470	3318	19375	43909	3067	405	889	3040	1541
$BERTbek_{Wiki}$	93.48	72.49	65.43	96.36	92.68	38.53	92.00	**60.79**	**61.50**	40.87
$BERTbek_{News-Small}$	**94.54**	**76.48**	**67.56**	**97.17**	**93.36**	**49.47**	**92.37**	60.50	60.68	40.98
$mBERT$	93.49	74.36	64.64	96.13	92.59	47.35	91.13	48.72	56.57	**42.16**

Table 5: Topic classification evaluation results for BERTbek and baseline models. F-score (F1), precision (Prec.) and recall (Rec.) scores are reported, best scores for each metric are highlighted.

Model Name	F1	Prec.	Rec.
$LR_{Word-ngrams}$	60.32	75.81	54.01
$LR_{Character-ngrams}$	66.33	**76.43**	58.59
$LR_{Word+Char-ngrams}$	68.69	76.36	62.42
RNN	70.81	72.60	69.11
$RNN_{Word-embeddings}$	71.88	75.23	68.81
CNN	68.41	63.98	71.86
$RNN + CNN$	72.77	76.08	69.74
$mBERT$	70.72	72.46	70.01
$BERTbek_{Wiki}$	71.41	75.08	70.00
$BERTbek_{News-Small}$	**73.31**	75.34	**72.31**

Table 6: NER performance results on the test set (F1 scores) for BERTbek and the baseline models. The highest score in each metric is highlighted.

Model	F1
$LSTM_{Word}$	59.08
$LSTM_{Char+Word}$	70.18
$LSTM_{Char+Word} + W.emb.$	74.41
$LSTM_{Char+Word} + W.emb. + CRF$	71.87
$mBERT$	75.14
$BERTbek_{Wiki}$	73.85
$BERTbek_{News-Small}$	76.88
$BERTbek_{News-big}$	**78.69**

data size is as crucial as the quality of data, if not more.

Training steps. Initially, all three BERTbek models were trained for 3M steps (as explained in Section 3.3.2). We further continued $BERTbek_{News-Big}$ model training until 6M steps to assess the model's performance gain. The model's performance over all the evaluation tasks keeps improving gradually for the first 3M steps, then it either starts to decline, or fluctuate around the highest score gained in the first 3M steps, indicating that training the BERT models more is not only time-consuming, but also does not necessarily gain any performance after all.

6. Conclusions and Future Work

In this work we presented BERTbek, consisting of three BERT pretrained language models for Uzbek, trained on different sizes and sources of text. We highlighted the process of obtaining a pretrained LM for a low resource language, such as data collection, tokenization, pretraining, in the example of the Uzbek language. Moreover, the resulting models were evaluated using three downstream NLP tasks, namely sentiment analysis, topic classification, and named entity recognition. The evaluation results showed that our BERTbek models outperformed all other baseline models in all three tasks, becoming state-of-the-art. Regardless of the relatively small size of the texts that were used to train our models, BERTbek has outperformed its multilingual counterpart (mBERT). The analysis results once more proved the statements from previous work that it is not only the bigger size of training data that increases BERT model's performance, but also the quality of the text that makes a big impact (Li et al., 2019), such as the cleanliness and structural diversity of the sentences in a corpus.

As a future work, following the trend of other ideas around pretraining BERT models for morphologically rich languages, especially with highly inflectional syntax, we aim to create morphologically-aware BERT language models for Uzbek as well as other similar languages in the Turkic family by using a tokenizer that splits words into chunks based on their prefix, stem, and suffixes, which will hopefully improve performance.

Furthermore, following the trend of multilingual BERT and other LMs, there is a plan to pretrain a multilingual BERT model including only strongly-related languages in the same language family (like multi-Turkic-BERT) to analyse the performance differences from multilingual BERT itself, as well as their monolingual counterparts in various NLP tasks, both in multilingual and monolingual evaluation settings. It would be interesting for truly-low-resource languages in the family, such as Turkmen and Karakalpak, where available raw text is not even enough for pretraining monolingual LMs,

to see if they profit from gained knowledge from resource-rich languages in the same family, such as Turkish.

7. Data Availability

All the code used in this work are openly available at https://github.com/elmurod1202/BERTbek. Also, the BERTbek models have been uploaded to the HuggingFace Models Hub at https://huggingface.co/elmurod1202/bertbek-news-big-cased.

8. Conflicts of Interest

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

9. Bibliographical References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. *arXiv preprint arXiv:2004.00033*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

Bobur Allaberdiev, Gayrat Matlatipov, Elmurod Kuriyozov, and Zafar Rakhmonov. 2024. Parallel texts dataset for uzbek-kazakh machine translation. *Data in Brief*, pages 110–194.

M. Aminov, B. Ahmedov, H. Boboev, T. Daminov, T. Dolimov, T. Jo'raev, A. Ziyo, N. Ibrohimov, N. Karimov, H. Karomatov, N. Komilov, A. Mansur, J. Musaev, E. Nabiev, A. Oripov, T. Risqiev, N. Tuxliev, D. Shorahmedov, R. Shog'ulomov, T. Qo'ziev, S. G'ulomov, , and A. Hojiev. 2000-2006. *O'zbekiston milliy ensklopediyasi*. "Ozbekiston milliy ensiklopediyasi" Davlat ilmiy nashryoti, Tashkent, Uzbekistan.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Vít Baisa, Vít Suchomel, et al. 2012. Large corpora for Turkic languages and unsupervised morphological analysis. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA)*.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Roberta Rodrigues de Lima, Anita MR Fernandes, James Roberto Bombasar, Bruno Alves Da Silva, Paul Crocker, and Valderi Reis Quietinho Leithardt. 2022. An empirical comparison of portuguese and multilingual bert models for auto-classification of ncm codes in international trade. *Big Data and Cognitive Computing*, 6(1):8.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Labehat Kryeziu and Visar Shehu. 2022. A survey of using unsupervised learning techniques in building masked language models for low resource languages. In *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–6. IEEE.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Elmurod Kuriyozov, Yerai Doval, and Carlos Gomez-Rodriguez. 2020. Cross-lingual word embeddings for turkic languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4054–4062.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2022. Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek. *Lecture Notes in Artificial Intelligence*, 13212:232–243.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *In Proceedings of the International Conference of Learning Representations (ICLR 2020).*

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, Hong Yu, et al. 2019. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, 7(3):e14830.

Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. 2015. Component-enhanced chinese character embeddings. *arXiv preprint arXiv:1508.06669*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: getting inside bert's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Khabibulla Madatov, Shukurla Bekchanov, and Jernej Vičič. 2022. Accuracy of the uzbek stop words detection: a case study on "school corpus". *CEUR Workshop Proceedings*, 3315:107 – 115.

Khabibulla Madatov, Shukurla Bekchanov, and Jernej Vičič. 2023. Automatic detection of stop words for texts in uzbek language. *Informatica*, 47(2).

B Mansurov and A Mansurov. 2021a. Uzbek cyrillic-latin-cyrillic machine transliteration. *arXiv preprint arXiv:2101.05162*.

B Mansurov and A Mansurov. 2021b. Uzbert: pretraining a bert model for uzbek. *arXiv preprint arXiv:2108.09814*.

Gayrat Matlatipov and Zygmunt Vetulani. 2009. Representation of Uzbek morphology in prolog. In *Aspects of Natural Language Processing*, pages 83–110, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sanatbek Matlatipov, Hulkar Rahimboeva, Jaloliddin Rajabov, and Elmurod Kuriyozov. 2022. Uzbek sentiment analysis based on local restaurant reviews. *CEUR Workshop Proceedings*, 3315:126 – 136.

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. In *EMNLP-IJCNLP 2019*, pages 6256–6261, Hong Kong, China. Association for Computational Linguistics.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and bert models for maltese. *arXiv preprint arXiv:2205.10517*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NACL 2018*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Ilyos Rabbimov, Iosif Mporas, Vasiliki Simaki, and Sami Kobilov. 2020. Investigating the effect of emoji in opinion classification of uzbek movie review comments. In *International Conference on Speech and Computer*, pages 435–445. Springer.

IM Rabbimov and SS Kobilov. 2020. Multi-class text classification of uzbek news articles using machine learning. In *Journal of Physics: Conference Series*, volume 1546.1, page 012097. IOP Publishing.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059.

Ulugbek Salaev, Elmurod Kuriyozov, and Carlos Gómez-Rodríguez. 2022a. A machine transliteration tool between uzbek alphabets. *CEUR Workshop Proceedings*, 3315:42 – 50.

Ulugbek Salaev, Elmurod Kuriyozov, and Carlos Gómez-Rodríguez. 2022b. Simreluz: Similarity and relatedness scores as a semantic evaluation dataset for uzbek language. *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, page 199 – 206.

Ulugbek I. Salaev, Elmurod R. Kuriyozov, and Gayrat R. Matlatipov. 2023. Design and implementation of a tool for extracting uzbek syllables. *Proceedings of the 2023 IEEE 16th International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering, APEIE 2023*, page 1750 – 1755.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maksud Sharipov, Jamolbek Mattiev, Jasur Sobirov, and Rustam Baltayev. 2022. Creating a morphological and syntactic tagged corpus for the uzbek language. *CEUR Workshop Proceedings*, 3315:93 – 98.

Maksud Sharipov and Ollabergan Yuldashov. 2022. Uzbekstemmer: Development of a rule-based stemming algorithm for uzbek language. *CEUR Workshop Proceedings*, 3315:137 – 144.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert–czech bert-like model for language representation. *arXiv preprint arXiv:2103.13031*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models

for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

TBA. Uzner: Named entity recognition dataset and its analysis for uzbek language. Submitted for a review around the same time.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical bert models for sequence labeling. *arXiv preprint arXiv:1909.00100*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Vilares, Marcos Garcia, and Carlos Gómez-Rodríguez. 2021. Bertinho: Galician bert representations. *arXiv preprint arXiv:2103.13799*.

David Vilares, Michalina Strzyz, Anders Søgaard, and Carlos Gómez-Rodríguez. 2020. Parsing as pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9114–9121.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y Ng. 2016. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*.

Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Linting Xue and al. et. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.