On Deep Classification of Low-Resource Turkic Languages

Michael Miller
University of Central Florida
Orlando, Florida, USA

Abstract

This project investigates text classification in low-resource Turkic languages, with a focus on Uzbek. To address the lack of significant parallel corpora, we translated a Russian news dataset into Uzbek using Facebook's `M2M_100` Neural Machine Translation (NMT) model. The resulting translated corpus was prepared for training a Very Deep Convolutional Neural Network (VDCNN) classifier by combining it with an existing Uzbek news classification dataset. Critical failure modes in multilingual NMT, such as repetition, hallucination, and context loss, are examined. A naive cleanup architecture significantly improved text readability, though more than 20% of translations remained semantically irrecoverable. While originally intended to measure low-resource classification performance, this project steered towards the present limitations of many-language NMT models in truly low-resource settings. We set the stage for future work in classification of underrepresented languages in our code repository, as our full evaluation pipeline including BLEU scoring, chrF++ scoring, and classifier training was hindered by the technical limitations of Google Colab.

Code is available at https://github.com/Dagomara/Uzbek-VDCNN.

1 Introduction & Problem Statement

In the context of low-resource languages, text classification is difficult because of limited classified corpora. Even though enough data may exist for a low-resource language to create Neural Machine Translation (NMT) embeddings between the target language and a higher-resource language such as English or Chinese, there may not be enough tagged data to train a classifier effectively on the low-resource language. Chiefly, there is not enough data to adequately train a transformer-based model on this language, though meta-learning or transfer learning may help [14].

As one of the first languages to be translated by machine, Russian is not a low-resource language by most standards. However, the Turkic languages such as Tatar, Uzbek, Turkish, etc. are all very low-resource compared to Russian and English, with substantial text classification datasets only appearing in the past three years [18] [2] [4] [1]. There are limited texts available for these languages, so Transformer-based learning is still error-prone. This is partially because Uzbek is a low-resource language, but is also because Uzbek is typically dominated by Russian translation, just like the other Turkic languages. Due to the political situation in those regions of the world, these languages are all typically translated into Russian instead of other languages (including each other). For foreign onlookers, it is difficult to get accurate translations of news articles or songs in these low-resource languages due to the limited training data models have seen bridging English and Uzbek.

Thankfully, smaller architectures such as RNN and CNN can provide decent categorization efficiency with low resource overhead and substantially-lower amounts of training data than what transformers require [18]. Still, there is a need for more training data on these low-resource Turkic languages. While we will focus on work in the Uzbek language, results have been seen to generalize across the Turkic languages listed above [17].

Because of geographical relations, there is a lot of bilingual English-Turkish text available, making English-Turkish MT models and embeddings recently competitive. Even though Turkish could be seen as a low-resource language in itself, English-Turkish translation tasks see high enough scores to use Turkish as a baseline for NMT (BLEU: 34.6) [1] between English and Turkish, eventually into Uzbek.

2 Related Work

To classify Uzbek language texts, corpora is needed. The Russian Social Media Text Classification Kaggle dataset (RSMTCKd) [3] is an example dataset for classification of a high-resource interstitial language to translate to Uzbek. A recent work from Kuriyozov *et al* [18] provides an excellent starting point for any classifiers trained in this experiment. The Kuriyozov dataset includes over 120 million Uzbek words of news article text, classified into the categories seen in Table 1.

tk introduce bleu and other metrics. The work of Khusainov *et al* [17] employs parallel training data in the languages of Tatar, Kazakh, Kyrgyz, Crimean Tatar, Uzbek, Turkish, and Russian to build a Tatar-Russian language model which can then map Tatar to the other 5 Turkic languages mentioned. This technique generated state-of-the-art NMT results for all language pairs permutated between the options. The work references Uzbek as the greatest in need of more training data. This work promoted increases in BLEU score, which focuses on n-gram precision between translations. Another popular metric for measuring translation quality is chrf++, which focuses on n-gram F-score with character matches.

Since Khusainov's work, some multi-language models have started offering Uzbek support. Chiefly, Microsoft's Marian NMT model provides Russian-Northern Uzbek support out of the box [16], and the Technical University of Darmstadt's UKP Lab hosts an *EasyNMT* library, offering [6] ru-uz translation through Facebook's `M2M_100` model [12]. Curiously, Facebook Research's No Language Left Behind [11] offers translation pairs including a high-scoring English-Uzbek translation model, but not a Russian-Uzbek pathway, implying Russian corpora would need to be translated to English as a middle ground if that model were used. Uzbek NMT has been verified through other models such as BERTbek [19], Turkic-Russian translation through Tatar encodings [17], and Transformer architectures [21] [6] [12] [11] as already discussed.

NLP researcher Max K's *rus_news_classifier* dataset [7] includes a news dataset with similar classifications to Kuriyozov *et al*'s dataset for classification, as seen in 1. It contains over 70k sequences of news text ranging from 18 to 3.55k characters on average, with some outliers extending to 35.3k characters.

[1] https://www.promptlayer.com/models/opus-mt-en-trk-b0a5

Conneau *et al*'s 2017 *Very Deep Convolutional Networks for Text Classification* [8] improved over state-of-the-art non-transformer text classification tasks through a character-level examination of text data, similar to how CNNs are used in Computer Vision. It holds relevance today in the low-resource language space as a transformer-based architecture requires significantly more data to generalize character-level patterns, such as typo understanding.

Table 1: Comparison of Dataset Categories

Kuriyozov	rus_news_classifier
Local (Mahalliy)	climate (климат)
World (Dunyo)	conflicts (конфликты)
Sport (Sport)	culture (культура)
Society (Jamiyat)	economy (экономика)
Law (Qonunchilik)	gloss (глянец)
Tech (Texnologiya)	health (здоровье)
Culture (Madaniyat)	politics (политика)
Politics (Siyosat)	science (наука)
Economics (Iqtisodiyot)	society (общество)
Auto (Avto)	sports (спорт)
Health (Salomatlik)	travel (путешествия)
Crime (Jinoyat)	
Photo (Foto)	
Women (Ayollar)	
Culinary (Pazandachilik)	

3 Methodology

The goal for this paper is to train a Very Deep Convolutional Network (VDCNN) for text classification on Uzbek text using NMT-translated datasets from the Russian Language. Our pipeline is simple and illustrated as follows:

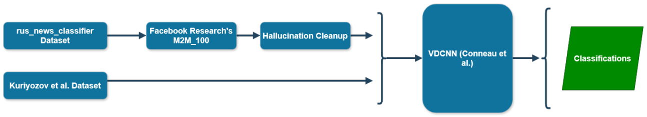

Figure 1: Pipeline for the model.

3.1 NMT Translation to Target Language

In addition to Kuriyozov *et al*'s dataset, we will be selecting the *rus_news_classifier* dataset for machine translation into Uzbek. Among other available corpora, we posit this choice provides the best starting point to estimate NMT dataset quality between Russian and Uzbek, due to its size and similarity to the *Kuriyozov* dataset. Kuriyozov *et al*'s dataset has five categories (culture, economy, health, sport, politics) in common with the *rus_news_classifier* dataset, and adds six additional categories, bolstering training size for a neural classifier when the two form a superset.

The Russian dataset is stripped of outlier-length entries and machine translated into Uzbek using Facebook Research's `M2M_100` model. Due to resource limitations, we were limited to a sample of 1600 total translated entries. Due to difficulties described in 4, the translations faced many issues and required further cleanup and dehallucination. Some of this was a manual process, but a large amount of it was automated. These translations are evaluated with BLEU and chrf++ scores to ensure quality and accuracy of sequences produced. Similar to *TAPE* (Text Attack and Perturbation Evaluation) few-shot Russian language understanding benchmark [5], tools such as ButterFingers, Emojify, EDA$_{delete}$ and EDA$_{swap}$ are used to make test perturbations of the new large Uzbek corpora.

3.2 Training of VDCNN to Classify Text

A VDCNN is used as recommended by Conneau *et al* to classify the Uzbek text corpora. VDCNN classification results are examined through mAP scores, Top-1 scores, and similar metrics. With a translated dataset of purposely-imperfect text entries, this model is forced to not overfit data and instead learn to generalize across novel spellings and contexts of words. Of course, it will require more training data to overcome the typo and emoji hurdles, hence our original NMT procurement of additional text.

4 Evaluation and Results

The evaluation phase of this project was intended to include both translation quality metrics (BLEU and chrF++) and a downstream classification task using the translated data. Unfortunately, due to technical constraints encountered while working on Google Colab, e.g. persistent runtime errors, loss of GPU access, and loss of session data, it was not possible to complete these evaluations in time. Despite these limitations, the core goal of translating a dataset into Uzbek and preparing it for evaluation was completed, though with limited success. The translated outputs are available and ready for future evaluation when computational resources are more stable or available, though its use case is more as a comparative benchmark for future Russian-to-Uzbek translation models.

While quantitative results are not presented in this report, qualitative inspection of the translated text through Yandex Translate indicated that the system successfully handled a few standard constructions and some domain-specific terminology. Further work would include computing BLEU and chrF++ scores and assessing performance on the intended classification task, once infrastructure issues are resolved. Two native speakers were contacted, but neither responded in time for a user study.

Before the cleanup step, translation did see many hallucinatory issues. Regex-based and frequency heuristics methods were both leveraged for cleanup. Ultimately, linting was able to remove long suffix hallucination from 1231 of 1600 entries to fewer than 571 entries, and repeated word hallucinations were removed from 763 samples, though some slipped through the cracks.

For example,

Original Russian Text:

> Житель Москвы сходил на сеанс эротического массажа, после которого умер. Об этом сообщает Telegram-канал Mash. По информации издания, 31-летний москвич заказывал сеанс массажа с последующими интимными услугами в квартире на бульваре Яна Райниса.

English Translation:

> A resident of Moscow went to an erotic massage session, after which he died. This is reported by the Telegram channel

> Mash. According to the newspaper, the 31-year-old Muscovite ordered a massage session followed by intimate services in an apartment on Jan Reinis Boulevard.

Facebook-NMT Uzbek Result:

> Moskovski masajni Telegram kanal Mash. 31 yaşına Moskovich, Yan Reynis'nin bulvarda bir apartmanda bir masaj seansizga masajizdi. Bir süre sonra, bir adamın vücudu banyo'da buldum. U'qadi laklar polisdi, o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi o'qadi. Mosquita'nın ölümünün tamadi bilan bilan bilan bilan bilan bilan bilan bilan. Tshekni'nin 3 laklar, kimliklar bo'yadi, Orgiya'dan sonra Grozno'da kayboldi. Bir partiyada, kompaniyada 4 qoladi aktivni video filmdi, o'ziyiyiyi-yiyiyiyiyiyiyiyiyiyiy.

Linting output:

> Moskovski masajni Telegram kanal Mash. 31 yaşına Moskovich, Yan Reynis'nin bulvarda bir apartmanda bir masaj seansizga masajizdi. Bir süre sonra, bir adamın vücudu banyo'da buldum. U'qadi laklar polisdi, o'qadi. Mosquita'nın ölümünün tamadi bilan. Tshekni'nin 3 laklar, kimliklar bo'yadi, Orgiya'dan sonra Grozno'da kayboldi. Bir partiyada, kompaniyada 4 qoladi aktivni video filmdi, o'ziyi.

Even with repeated tokens removed, the result unfortunately translated in an unintelligible way and lost most semantic meaning. One could infer the original message only if they knew it before making their guess. Manual review confirmed major readability improvements, but most samples remained incoherent if major hallucination occurred during inference time.

5 Discussion & Challenges

Seq2Seq models give generic and unapplicable translations during greedy decoding because high-probability response tokens are safe but uninformative. Of all hallucinations, `M2M_100`'s largest struggle was with repetition. We posit this to be from their model overfitting on high-frequency dataset patterns. During initial evaluation, many repeating phrases and suffixes were seen across samples. The `M2M_100` model [9, 13, 22, 23] we used for translation was trained using CCAligned [10] and CCMatrix [24] datasets, as well as others. These papers do not mention Uzbek language, rather Turkish and some other Slavic languages, so the source of the Uzbek data is uncertain. As an LRL, Uzbek is difficult to obtain a diverse dataset for. Finding a substantial Russian-Uzbek translation pair dataset is an even more challenging task, one we could not do at scale for the scope of this project, so it is plausible that Facebook also struggled to find valid pairs without machine translating Russian corpora into Uzbek first. Said machine translation would introduce its own latent biases from whatever translation tools were used in said circumstances. Therefore, it is entirely possible that their reported Uzbek translation scores were biased towards their generated (or low-volume) dataset, making for a poor translation when unseen themes were introduced.

Considering Uzbek is an LRL, the Transformer-based architecture used could not comprehend less-common patterns and ended up repeating "safe" or "common" tokens ad-nauseum before escaping a loop. The repetition introduced a further problem: It is a known issue of Transformer models to experience issues in retaining long-term dependencies. By attending heavier to the latter tokens in a sequence as it translates, by the time the model escapes a repetition loop, it has lost its understanding of the source content it was intended to translate. In other words, with each looped suffix or phrase, less long-term dependencies of the source text sequence were retained, causing loss of information of the source text. This produced incoherent or otherwise unrelated text after the looped words. Standard Seq2Seq translation maximizes $P(T \mid S)$, which can give generic, contextually weak, irrelevant responses like we saw. The optimization techniques used for `M2M_100` [26] could be improved to Maximum Mutual Information (MMI) [20], which yields substantive gains in conversational and translation models alike [15, 25]. We see many examples in the cleaned dataset in which a translation has little to do with the source context. From my limited conversations with native Uzbek speakers, many of the translations which suffered from repetitive tokens escaped repetition only to start writing about unrelated concepts such as "Microsoft" or "Telegram channels," implying a very limited technology-focused parallel corpora was used by Facebook to train the Russian-Uzbek model.

Another possible influence of these hallucinations would be exposure bias during training, which is common in Seq2Seq models. This is where the model becomes reliant on its own predictions and spirals into repetition. Non-autoregressive models, retrieval-augmented MT, and language adapters/fine-tuned layers demonstrate broader options for budding LRL translation architectures without encountering that exposure bias as much. These hallucinated, misleading translations in sensitive domains (e.g., news, healthcare) can be dangerous from an ethical perspective, especially when NMT is used in high-stakes low-resource contexts. As a result, it is critical we navigate away from conditions which increase the likelihood of context derailment or other hallucinations we analyzed. This work highlights the fragility of large multilingual MT systems in true low-resource settings, especially when the training data is likely noisy or synthetic.

6 Concluding Remarks

This project set out to explore the challenges and opportunities of performing text classification on low-resource Turkic languages, with a specific focus on Uzbek. Although full evaluation of translation and classification performance was impeded by technical constraints, the process of assembling, translating, and preprocessing a cross-lingual dataset still uncovered several critical insights.

First, we demonstrated that even state-of-the-art multilingual models like Facebook's `M2M_100` struggle with stability and fidelity when tasked with Russian-to-Uzbek translation. Hallucinations, repetition loops, and context derailment were common, raising important concerns about the reliability of these models for downstream tasks in truly low-resource contexts.

Second, this work highlighted the fragility of low-resource pipelines where parallel corpora are limited, often synthetic, and potentially biased. While traditional evaluation metrics (e.g., BLEU, chrF++) could not be computed in time, the translation output itself offered a sample case study in the failure modes of these large multilingual NMT systems. These findings are useful to those seeking to deploy such models without extensive fine-tuning or native-language oversight, much like our team did.

Finally, while we were not able to train or evaluate the VDCNN classifier due to Colab resource instability, the architecture and dataset pairing remain viable. All of the code works, but training time ran over even our paid runtime limits. The cleaned, translated dataset is prepared and ready for classification experimentation when infrastructure permits. Beyond the scope of this course, "our team" is still interested in finishing the classification task.

This project affirms that robust multilingual NLP for low-resource languages remains an unsolved problem. It also affirms that honest reflection on translation artifacts, preprocessing tradeoffs, and data sourcing is just as critical as numerical benchmarks. Future work will continue to pair practical experimentation with a critical eye toward the data and model assumptions that underlie performance.

References

[1] [n. d.]. Bert Turkish Text Classification · Models · Dataloop. https://dataloop.ai/library/model/savasy_bert-turkish-text-classification/

[2] [n. d.]. gurkan08/bert-turkish-text-classification · Hugging Face. https://huggingface.co/gurkan08/bert-turkish-text-classification

[3] [n. d.]. Russian Social Media Text Classification. https://www.kaggle.com/datasets/mikhailma/russian-social-media-text-classification

[4] [n. d.]. savasy/bert-turkish-text-classification · Hugging Face. https://huggingface.co/savasy/bert-turkish-text-classification

[5] 2024. RussianNLP/tape · Datasets at Hugging Face. https://huggingface.co/datasets/RussianNLP/tape

[6] 2025. https://github.com/UKPLab/EasyNMT

[7] 2025. data-silence/fasttext-rus-news-classifier · Hugging Face. https://huggingface.co/data-silence/fasttext-rus-news-classifier [Online; accessed 25. Feb. 2025].

[8] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very Deep Convolutional Networks for Text Classification. arXiv:1606.01781 [cs.CL] https://arxiv.org/abs/1606.01781

[9] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2019. A Massive Collection of Cross-Lingual Web-Document Pairs. *arXiv preprint arXiv:1911.06154* (2019).

[10] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2020. CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs. arXiv:1911.06154 [cs.CL] https://arxiv.org/abs/1911.06154

[11] Facebook Research. [n. d.]. https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling

[12] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. arXiv:2010.11125 [cs.CL] https://arxiv.org/abs/2010.11125

[13] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. *arXiv preprint* (2020).

[14] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor O. K. Li. 2018. Meta-Learning for Low-Resource Neural Machine Translation. arXiv:1808.08437 [cs.CL] https://arxiv.org/abs/1808.08437

[15] Chenye Zhu Harrison Ho. 2017. Neural Conversational Model with Mutual Information Ranking. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761918.pdf. [Accessed 22-04-2025].

[16] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, Melbourne, Australia, 116–121. http://www.aclweb.org/anthology/P18-4020

[17] Aidar Khusainov, Dzhavdet Suleymanov, Rinat Gilmullin, and Ajrat Gatiatullin. 2018. Building the Tatar-Russian NMT System Based on Re-translation of Multilingual Data. In *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings* (Brno, Czech Republic). Springer-Verlag, Berlin, Heidelberg, 163–170. https://doi.org/10.1007/978-3-030-00794-2_17

[18] Elmurod Kuriyozov, Ulugbek Salaev, Sanatbek Matlatipov, and Gayrat Matlatipov. 2023. Text classification dataset and analysis for Uzbek language. arXiv:2302.14494 [cs.CL] https://arxiv.org/abs/2302.14494

[19] Elmurod Kuriyozov, David Vilares, and Carlos Gómez-Rodríguez. 2024. BERTbek: A Pretrained Language Model for Uzbek. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, Maite Melero, Sakriani Sakti, and Claudia Soria (Eds.). ELRA and ICCL, Torino, Italia, 33–44. https://aclanthology.org/2024.sigul-1.5/

[20] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). Association for Computational Linguistics, San Diego, California, 110–119. https://doi.org/10.18653/v1/N16-1014

[21] Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A Large-Scale Study of Machine Translation in Turkic Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5876–5890. https://doi.org/10.18653/v1/2021.emnlp-main.475

[22] Facebook Research. 2020. Beyond English-Centric Multilingual Machine Translation. https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100.

[23] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944* (2019).

[24] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. arXiv:1911.04944 [cs.CL] https://arxiv.org/abs/1911.04944

[25] Huitao Shen. 2019. Mutual Information Scaling and Expressive Power of Sequence Models. arXiv:1905.04271 [cs.LG] https://arxiv.org/abs/1905.04271

[26] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. arXiv:2008.00401 [cs.CL] https://arxiv.org/abs/2008.00401