proceedings

Proceedings

Building a New Sentiment Analysis Dataset for Uzbek Language and Creating Baseline Models †

Elmurod Kuriyozov [1,*] and Sanatbek Matlatipov [2]

[1] CITIC, Grupo LYS, Departamento de Computación. Facultade de Informática, Campus de Elviña, Universidade da Coruña, 15071 A Coruña, Spain

[2] Applied Mathematics and Computer Analysis Department, National University of Uzbekistan, University Str. 4, Tashkent 100174, Uzbekistan

* Correspondence: e.kuriyozov@udc.es; Tel.: +34-698-374-159

† Presented at the 2nd XoveTIC Congress, A Coruña, Spain, 5–6 September 2019.

Published: 2 August 2019

Abstract: Making natural language processing technologies available for low-resource languages is an important goal to improve the access to technology in their communities of speakers. In this paper, we provide the first annotated corpora for polarity classification for Uzbek language. Our methodology considers collecting a medium-size manually annotated dataset and a larger-size dataset automatically translated from existing resources. Then, we use these datasets to train sentiment analysis models on the Uzbek language, using both traditional machine learning techniques and recent deep learning models.

Keywords: sentiment analysis dataset; Uzbek language; sentiment classification; Natural Language Processing; deep learning

1. Introduction

The advancement of technologies in the field of Natural Language Processing (NLP) over the past few years has led to achieve very high accuracy results, allowing the creation of useful applications that play an important role in many areas now. In particular, the adoption of deep learning models has boosted accuracy figures across a wide range of NLP tasks. As a part of this trend, sentiment classification, a prominent example of the applications of NLP, has seen substantial gains in performance by using deep learning approaches compared to its predecessor approaches [1]. However, low-resource languages still lack access to those performance improvements due to the requirements of significant amounts of annotated training data to work well. The language we focus on in this paper is Uzbek, which is spoken by more than 33 million native speakers in Uzbekistan as well as neighbouring countries. Uzbek is a Turkic language that is the first official and only declared national language of Uzbekistan. Uzbek is a null-subject, highly agglutinative language [2].
The main contribution of this paper is the annotated dataset for sentiment analysis in Uzbek language, obtained from Google Play Store reviews and a larger dataset by automatically translating an existing English dataset. Furthermore, we define the baselines for sentiment analyses in Uzbek by considering both traditional machine learning methods as well as recent deep learning techniques fed with fastText pre-trained word embeddings [3]. Although all the tested models are relatively accurate and differences between models are small, the neural network models tested do not manage to substantially outperform traditional models.

2. Experiments & Results

For data collection, the list of top 100 applications from Google Play App Store used in Uzbekistan have been selected, retrieving their review texts and related star rating using Google Play Store API.

Collected text has been cleaned (Removing names, brands, tags, links and emojis) and the reviews in Cyrillic alphabet have been converted to Latin. For the annotation process, the main task was binary classification: to label the them as positive or negative, so two authors manually labeled the reviews. A third score was obtained from the star rating of the review itself: positive if a certain review has more than 3 stars, otherwise negative (Majority of 3-starred reviews had negative sentiment). Finally, the review was given a polarity according to the majority label. This process resulted into 2500 reviews annotated as positive and 1800 as negative, 4300 in total.

In order to further extend the resources to support sentiment analysis, another larger dataset was obtained through machine translation. An available English dataset of positive and negative reviews of Android apps, containing 10,000 reviews of each class, was automatically translated using MTRANSLATE, an unofficial Google Translate API from English to Uzbek. We manually went through the translation results quickly and examined a random subset of the reviews, large enough to make a reasonable decision on overall accuracy. Although the translation was not clear enough to use for daily purposes, the meaning of the sentences was preserved, and in particular, the sentiment polarity was kept (except for very few exceptional cases). As a result, we have obtained almost 20,000 translated reviews, balances between polarity classes. Both obtained datasets have been split into a training and a test set with a 90:10 ratio, for the experiments.

To introduce the baseline models for Uzbek sentiment analysis, we chose various classifiers from different families, including different methods of Logistic Regression (LR), Support Vector Machines (SVM), and recent Deep Learning methods, such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). The standard parameters as well as the performance metric were chosen for all methods [4]. We implemented LR and SVM models by means of the Scikit-Learn [5] machine learning library in Python with default configuration parameters. In the case of Deep Learning models, we used Keras [6] on top of TensorFlow [7].

Table 1 shows the classification accuracy obtained in three different configurations: a first one working on the manually annotated dataset (ManualTT), a second one on the translated dataset (TransTT) and a third one in which training was performed on translated dataset while testing was performed on the manually annotated dataset.

Table 1. Accuracy results with different training and test sets. **ManualTT**—Manually annotated Training and Test sets. **TransTT**—Translated Training and Test sets. **TTMT**—Translated dataset for Training, Annotated dataset for Test set.

Methods Used	ManualTT	TransTT	TTMT
Support-vector Machines based on linear kernel model	0.8002	0.8588	0.7756
Logistic Regression model based on word ngrams	0.8547	0.8810	0.7720
Recurrent + Convolutional neural network	0.8653	0.8864	0.7850
Recurrent Neural Network with fastText pre-trained word embeddings	0.8782	0.8832	0.7996
Logistic Regression model based on word and character ngram	0.8846	**0.8956**	**0.8145**
Recurrent Neural Network without pre-trained embeddings	0.8868	0.8832	0.8052
Logistic Regression model based on character ngrams	0.8868	0.8945	0.8021
Convolutional Neural Network (Multichannel)	**0.8888**	0.8832	0.8120

We achieved our best accuracy (89.56%) on the translated dataset using a logistic regression model using word and character n-grams. The modern deep learning approaches have shown very similar results, without substantially outperforming classic ones in accuracy as they tend to do when used for resource-rich languages.

Although the results obtained have been good in general terms, those obtained for deep learning models have not clearly surpassed the results obtained by traditional classifiers. This is mainly due to the highly agglutinative aspect of Uzbek language, which it harder to rely on word embeddings.

Funding: This work has received funding from the ANSWER-ASAP project (TIN2017-85160-C2-1-R) from MINECO, and from Xunta de Galicia (ED431B 2017/01 and Oportunius program). Elmurod Kuriyozov receives funding for his PhD from El-Yurt-Umidi Foundation under the Cabinet of Ministers of the Republic of Uzbekistan.

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.

References

1. Barnes, J.; Klinger, R.; Walde, S.S.I. Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets. *arXiv* **2017**, arXiv:1709.04219.
2. Matlatipov, G.; Vetulani, Z. Representation of Uzbek Morphology in Prolog. In *Aspects of Natural Language Processing*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5070.
3. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
4. Kuriyozov, E.; Matlatipov, S.; Alonso, M.A.; Gómez-Rodrıguez, C. Deep Learning vs. Classic Models on a New Uzbek Sentiment Analysis Dataset. In Proceedings of the Human Language Technologies as a Challenge for Computer Science and Linguistics—2019, Roznan, Poland, 6–8 November 2009; pp. 258–262.
5. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
6. Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 5 September 2019).
7. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: tensorflow.org. (accessed on 5 September 2019).

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).