Question Classification in Albanian Through Deep Learning Approaches

Evis Trandafili, Nelda Kote, Gjergj Plepi
Faculty of Information Technology
Polytechnic University of Tirana
Tirana, Albania

***Abstract*—In recent years, there is growing interest in intelligent conversation systems. In this context, Question Classification is an essential subtask in Question Answering systems that determines the question type, therefore, also the type of the answer. However, while there is abundant research for English, little research work has been carried out for other languages. In this paper we deal with classification of questions in the Albanian language which is considered a complex Indo-European language. We employ both machine learning and deep learning approaches on a large corpus in Albanian based on the six-class TREC dataset with approximately 5000 questions. Experiments with and without stop-words show that the impact of stop-words is significant in the accuracy of the classifier. Extensive comparison of algorithms for the task of question classification in Albanian show that deep learning algorithms outperform conventional machine learning approaches. To the best of our knowledge this is the first approach in literature for classifying questions in Albanian and the results are highly comparable to English.**

Keywords*—Question classification; deep learning; BiLSTM; transformer; RoBERTa; Albanian corpus; natural language processing

I. Introduction

Large amounts of unstructured text data are generated every day resulting in a continuously increasing demand for information retrieval. Intelligent conversational agents that use Artificial Intelligence (AI) algorithms can facilitate user interaction with computer systems, conduct a conversation and answer natural language questions. In this context, Natural Language Processing (NLP), as a branch of AI, deals with automatic comprehension, interpretation and manipulation of natural language. However, natural language is complex and implies many different forms of communication facets leading to the need to interpret not only narrative or confirmative phrases but also questioning ones. Therefore, question answering systems go beyond the simple retrieval of relevant documents and aim at generating answers in natural language [1], [2].

Question Classification (QC) is the core component of a QA system that directly affects the retrieved or generated answer. QC deals with the assignment of question labels based on the corresponding answer type [3], [4]. This process is usually modelled as a text classification problem employing Machine Learning and Deep learning approaches. A lot of research works has been done to develop intelligent systems for QC in different languages and the recent performance of state of the art QA systems is impressive. However, QA systems lack in performance on low resource languages and create linguistic barriers that hinder the free flow of knowledge, business and communication.

In this paper we deal with question classification for a low resource language like Albanian through modelling this task as a classification problem. An Albanian corpus composed of approximately 5000 questions based on Text REtrieval Conference (TREC) Question Classification dataset with 6 classes of questions has been created. This corpus has been used with three traditional machine learning algorithms Support Vector Machines, Random Forest and Logistic Regression. For each classification algorithm we have explored the use of different approaches to extract information from raw text data, the classical approach tf-idf, FastText embeddings and the RoBERTa pre-trained language model. In addition, the corpus has been used in other experiments that employe three deep learning models, BiLSTM with and without attention, Transformer and RoBERTa. The overall evaluation of the models has been assessed using accuracy, precision, recall and F-score.

The experimental results show that stop-words have a high impact on the classification accuracy. Experimental comparison of different models shows that deep learning algorithms outperform the traditional machine learning approaches. Finally, although a first attempt for Albanian, our approach is comparable with the same approaches for English, leading to promising results for building question answering systems in Albanian and similar low resource languages.

The content of the paper is organized as follows: In Section II we cover Question Classification background and related works. In Section III we present the technologies and software libraries that we used in our experimental environment. In Section IV the Albanian dataset is introduced and its structure is explained in detail. Moreover, in Sections V and VI we introduce learning models, experimental design and analyze the results. Finally, we conclude with Section VII.

II. Related Work

There has been relatively little research done in the field of question classification tasks in the Albanian language. The first paper in this field treated this task as a traditional classification problem. The Albanian question corpus was used to evaluate the performance of three question classification models, which utilized the Support Vector Machine (SVM), Logic Regression

(LR) and Random Forest (RF) algorithms. The SVM model achieved the highest accuracy of 75.7% using FastText, while the use of RoBERTa resulted in a slightly lower accuracy of 0.6% [5]. The question-answering system proposed for the Albanian language in [6] does not implement the question classification component. The system has three modules: the first module preprocesses and indexes the documents, the second module analyzes the question, and the third module retrieves passages and extracts answers.

The current state of the question classification task in the Albanian language is still in its early stages, and further research and development are needed to enhance the performance of classification models. Nonetheless, the available approaches show promising results. Therefore, we will focus on existing studies related to languages other than Albanian.

The three components of question answering systems are question analysis, information retrieval, and answer analysis. The initial step of the question analysis component is question classification, which aims to determine the type of question and, consequently, the type of answer required. The question analysis component and its question classification step play an important role in the question answering systems as it can identify the correct class of a question and as a result restrict the possible answers [3, 4].

In question answering systems, the question type is a key factor that determines which type of questions a system can answer. These systems can be factoid, which answer questions related to facts, or non-factoid, which answer questions like mathematical calculations [7].

Previous work on the question classification problem can be analyzed from different aspects, such as the question classification categories, classification methods and algorithms, features, and so on. The answer to a question depends on the type of question, and questions can be classified into a category based on their common linguistic properties. Furthermore, their answers can also share common linguistic properties [8].

Numerous research studies have focused on developing artificial intelligence systems for question classification in various languages, including English [9], German [10], Italian [11], French [12], Spanish [13], and others.

The TREC [14] dataset, a popular resource for question classification tasks, is widely used in English and has also been translated into various other languages. This dataset contains a large collection of labeled questions and has been instrumental in the development and evaluation of many questions classification models. It is a dataset of 6,000 English questions, categorized into a main classification schema with six labels (human, description, abbreviation, entity, location, and numeric value,) and a fine-grained schema for each of the main categories. In [15], two classes (yes-no-explain and list) were added to the main taxonomy to improve question classification accuracy. The "yes-no-explain" class is used for questions that require a yes or no answer with an accompanying explanation. On the other hand, the "list" class is used for questions that have answers from a predefined list. An additional 250 questions were also added to the dataset. In [16], a Portuguese version of the TREC corpus was proposed.

The DISEQuA Corpus [17], a multilingual question answering corpus that uses a 7-class label classification schema (location, date, object, measure, person, organization, and other) is another widely used corpus. Additionally, paper [18] proposes a much larger taxonomy of 180 classes, making it one of the biggest taxonomies used in QA systems for question classification.

The SVM algorithm is frequently utilized in question classification tasks. In [19], an Arabic version of the TREC dataset is introduced and employed to train a two-stage classification model. The model involves using either the SVM, RF, or ME algorithm, followed by a CNN neural network. The machine learning algorithm is utilized to predict the primary class of the question in the first phase, and then the CNN is used in the second phase to predict the subclass. The SVM algorithm with CNN achieved the best performance, with an accuracy of 89%.

The use of top-words and dependency relations as features with the SVM algorithm in paper [20] resulted in an accuracy of 93.4% in the question classification task using the TREC dataset. Furthermore, the proposed solution in [21] to use syntactic and semantic analysis with SVM algorithms improves the performance of the model compared to the state-of-the-art models. The proposed hybrid method in [22] using the SVM algorithms with semantic and lexical feature extraction results in an accuracy of 96% on the TREC dataset.

The grammatical-based framework proposed in [23] performs better when using the J48 algorithm than SVM in factoid question classification approach, achieving an accuracy of 95.8%. The three main features of this approach are grammatical features, domain-specific features, and patterns.

In [24], the author proposed a Question Classification approach for Italian based on word embedding with sub-word information and Convolutional Neural Networks. The proposed approach is tested on the TREC dataset of questions in English and the Italian translated version, by using advanced vectors learned in an unsupervised manner using the skipgram model and character-based information to initialize the word embeddings. The Italian model achieved an accuracy of 80.42% using FastText, which is comparable to the English model accuracy of 80.16%.

In [25], the author proposed two models based on LSTM networks for question classification. These models were trained and evaluated on the heterogenous TREC and USC dataset, which features a two-level hierarchy annotation schema. The first model was used solely for predicting the main class of the question, while the second model added sub-class prediction to the question classification. The models achieved high accuracy values of 91.20% for the main class and 82.20% for the sub-class using 1000 h dimensions. The authors concluded that these types of networks are highly effective in the question classification task.

In [26], the author proposed the use of a BERT-based model for pregnancy question classification. Two attention mechanisms were evaluated: an additional layer on top of

BERT and the built-in self-attention mechanism of BERT, which resulted in better classification accuracy than the traditional models. The model that used an additional layer on top of BERT achieved the highest accuracy of 88%.

The new approach, which leverages data augmentation to create extra training instances, is presented in [27]. The effectiveness of this approach is evaluated on the TREC question classification datasets using pre-trained models such as BERT, RoBERTa, and DeBERTa. The results indicate that the need for labeled instances is reduced by up to 81.7%, achieving a new state-of-the-art classification accuracy of 98.11% on the TREC dataset.

III. Technologies and Software Libraries

Learning models used in this paper are trained and tested using Python programming language. Python is the most preferred language for computer scientists and researchers working in the field of NLP because of the versatility to use API-s and libraries that enhance performance growth [28].

As a programming environment we used the Jupyter Notebook. It is a free, open source, web based computational notebook that supports over 100 programming languages even if its name is a reference of three core programming languages, Julia, Python and R. Python is the most popular programing language with Jupyter notebook [29]. The notebook has two main components: The first component consists of front-end cells that hold text or code and are executed independently. This is very important for time consuming operations like dataset manipulations. The second component is the back-end kernel that executes the code inside the cells. Kernels can also be run on remote servers.

For our experimental environment, we installed Visual Studio IDE which fully supports work in Jupyter Notebook and Python. In the subsections below we present libraries and modules imported in Jupyter Notebook.

A. Pandas

Pandas is a mature Python library with a stable API used for data analysis [30]. Pandas library is used to deal with the .csv files where the dataset with questions is kept. Pandas offers build in functions that facilitate text manipulation. We used DataFrame, a two-dimensional data structure, to create the .csv file that contains both, questions in English, their equivalent in Albanian and the respective question tag.

B. Scikit-learn (Sklearn)

Ease of use and its computational efficiency make Scikit-learn (Sklearn) the most used Python library for supervised and unsupervised Machine Learning algorithms. It's task-oriented interface enables easy comparisons of different machine learning algorithms for a given dataset [31].

In our Question Classification task, we used the implementation of Support Vector Machine (SVM), Random Forest and Logistic Regression from Scikit-learn to build Question Classification models [5]. The main object in our supervised learning is the Scikit-learn estimator that implements a *fit* method with two arguments: an array with data and an array with labels. After calculating model learned parameters in *fit* method, we use the *predict* method with a single question input to predict/classify the question. By using the *score* method of the estimator, we compute the accuracy by default.

C. PyTorch

PyTorch is a Python framework which is very popular in deep learning research community. It assembles usability together with performance. PyTorch has proved its overall speed on several common benchmarks. It supports code as a model, offers an easy debugging, is efficient with other libraries and supports hardware accelerators [32]. PyTorch provides an array-based programming model accelerated by GPUs. NLP applications are successfully used with PyTorch because of the tensors which allow the use of GPUs to perform complex computational calculations with significantly improved performance.

D. Hugging-Face

Hugging Face Transformer is a neural network architecture by Google Brain based on attention mechanisms to draw global dependencies between input and output. It outperformed traditional encoder-decoder architectures in translation tasks [33]. Hugging Face is a great source of pre-trained machine learning models based on several transformer architectures like BERT [34], RoBERTa [35], BART [36], etc.

IV. Dataset Preparation

The Albanian questions dataset used in this paper is built by translating into Albanian the six-class fact-based questions dataset from Text REtrieval Conference (TREC) [14], originally with 5452 training examples and 500 test examples. The six-class tags used to identify the type of answer are shown in the Table I below.

TABLE I. Six-Class Tags in Questions Dataset

Question tag	Abbreviation	Description
0	DESC	Description and abstract concepts
1	ENTY	Entities
2	ABBR	Abbreviation
3	HUM	Human beings
4	NUM	Numeric values
5	LOC	Locations

Raw translation from Albanian language was achieved using the library translate-api in python. Google Translate is the most used online translation service, therefore we used its interface for translating the dataset from English to Albanian. However, taking in consideration that Albanian language has a limited translation support, after the automatic translation we checked the dataset manually. Two main issues were noticed:

- The translation was not very accurate.
- Some questions lost their relations with the respective tags.

Therefore, we revised the dataset manually by correcting translations and by removing questions considered not

appropriate for use in the dataset. As a result, the final dataset is composed of 4694 questions in Albanian. In Table II we show several examples of questions translated from English using the automatic translation.

The numbers of questions in each of the six classes in questions' dataset is shown in Fig. 1. We can easily observe that the class ABBR is highly unbalanced as the number of questions classified as abbreviations is merely 8% of the number of questions in other classes.

The problem of class imbalance is very popular in real world classification datasets. Questions asking for explanation of abbreviations are infrequent in our everyday life, as a result they are rarely found in written texts compared to other types of questions. This is clearly reflected even in questions' dataset. In this situation the classification risks to ignore the ABBR class.

The solution that we used to learn from the unbalanced questions dataset is the data augmentation technique. We manually increased the number of sentences in the ABBR class by creating new sentences through exchanging some words of the question with their synonyms.

TABLE II. EXAMPLES OF TRANSLATED QUESTIONS

Question tag	English Question	Albanian Question
DESC	How do you match a name to a social security number?	Si shoqërohet një emër me një numër të sigurimeve shoqërore?
ENTY	What's the only work by Michelangelo that bears his signature?	Cila është e vetmja vepër nga Michelangelo që mban firmën e tij?
ABBR	What is the abbreviation of the company name "General Motors"?	Cili është shkurtimi i emrit të kompanisë "General Motors"?
HUM	Who killed Gandhi ?	Kush e vrau Gandin?
NUM	How long does it take for your blood to make one complete trip through the body?	Sa kohë duhet për gjakun tuaj të bëjë një udhëtim të plotë përmes trupit?
LOC	What country is the world 's largest importer of cognac?	Cili vend është importuesi më i madh në botë i konjakut?

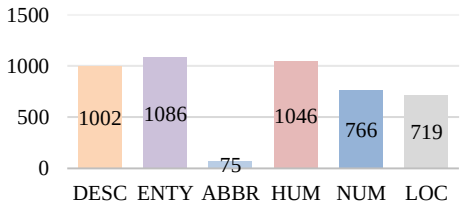

Fig. 1. Class distribution of questions.

Moreover, the most recommended pre-processing method that showed improvement in text classification tasks is stop-word removal. Stop-words are a set of commonly used words in a natural language that hold unimportant information in NLP applications. Their removal allows the model to learn from the most important words. However, the authors in [37] showed that stop-word removal does not always improve accuracy. Text Classification performance should always be fine-tuned by performing evaluations and determining the best combinations of pre-processing methods. As a pre-processing step, we used stop-word removal. We utilized the Spark-NLP [38] python library, which supports stop-word removal for Albanian language. A total of 223 frequently used words with low level information in Albanian language are removed from the dataset. Some examples of stop-words in Albanian are: "*ai*", "*ajo*", "*çfarë*", "*cili*", "*dhe*", "*unë*", etc. To test if stop-word removal is beneficial in terms of accuracy for our question classification model, we preserve a copy of the dataset, prior to stop word removal. The latter is also used for training the model and its accuracy is compared with the accuracy from models trained with the pre-processed dataset. We explore the effects of pre-processing with stop-words in Experiment 2 in Section VI.

In conventional Machine Learning algorithms, the words in the dataset are represented as numeric vectors or word embeddings in order for them to be fed as input to the models. In this paper we applied FastText [39], a popular neural network model used to generate word embeddings by splitting the word in n-grams. Due to its generalization capabilities and the ability to compute word embeddings for out-of-vocabulary words, FastText has shown good accuracy for small datasets and for low resource languages. In [5] we used tf-idf, FastText and RoBERTa to represent word embedding fed to Machine Learning models and the results showed that FastText outperformed in terms of accuracy.

V. LEARNING MODELS AND EXPERIMENTAL SETUP

We model the question classification task as a classification problem. If we are given a question u composed by a sequence of words $x = \{x_1, \cdots, x_n\}$, the classification model should predict the label y, by computing the probability $p\ (y \mid x)$. To effectively deal with this problem, we compared the accuracy of several classification algorithms divided in two main categories:

a) Classical machine learning algorithms: Support Vector Machines, Random Forest and Logistic Regression with tf-idf, FastText and RoBERTa as vector representations. We used the results obtained in [5].

b) Deep learning algorithms: *BiLSTM* [40] is a bidirectional LSTM which takes as input a sequence of vector representations for each word in the question/sentence and outputs the class/label of the question.

Transformer [33] is a new and powerful neural networks architecture developed by Google. In the classification context, it utilizes a similar approach with BiLSTM but using a different and more powerful architecture.

RoBERTa [35] is a transformer model pre-trained on large volumes of raw text data. For the Albanian language it is pre-trained using Wikipedia corpus. In our work, we used the RoBERTa pretrained Albanian tokenizer. RoBERTa transformer model takes as input a sequence of indexes from sentence words acquired from the tokenizer and outputs an array of base size 768 for each part of the sequence. The arrays

are then unified, and the average value is calculated and fed to a feed-forward network with two layers and an activation function, in our case RELU.

Moreover, the deep learning models that we used for training have different numbers of learnable parameters, frequently used to measure the classification performance. In Table III, we show the number of parameters in each model.

TABLE III. NUMBER OF PARAMETERS FOR EACH MODEL

Model	No of Parameters
BiLSTM	3.8 M
BiLSTM with attention	4.8 M
Transformer	4 M
RoBERTa	83.7 M

RoBERTa is the model with the highest number of parameters, 83.7M, while BiLSTM is the model with the least number of parameters. If we assess sequence-to-sequence models, BiLSTM with attention has the highest number of parameters while Transformer and BiLSTM have comparable number of parameters.

Classical Machine Learning and Deep Learning models are trained on a computer with a 16 cores CPU and a high-performing graphics processing unit, GPU NVIDIA A100-PCIE 40GB.

The classical Machine Learning algorithms were evaluated using 5-fold cross validation. Each model spent approximately 1-2 minutes to build.

On the other hand, to build deep learning models we performed the 70/10/20 train/validation/test splitting technique. In Table IV the distribution of classes during the split is shown.

TABLE IV. OVERVIEW OF TRAIN/VALIDATION/TEST

Split	Total	DESC	ENTY	ABBR	HUM	NUM	LOC
Train	3191	691	739	51	706	540	464
Validation	564	116	131	6	133	85	93
Test	939	195	216	18	207	141	162

To avoid overfitting, an early stopping method is implemented in the validation set. The models were fine-tuned for 100 epochs and a batch size of 64. Furthermore, we applied the Adam optimizer [41] as the most recommended optimizer in reducing loss and improving accuracy. When training BiLSTM, BiLSTM with attention and Transformer models, we applied $1e-3$ learning rate. Whereas, for RoBERTa model we used $1e-4$ learning rate [35]. Each of deep learning models spent 15-20 minutes when trained in the beforementioned computer hardware. Deep learning models require powerful processing units in order to ensure efficiency and reduce time consumption.

VI. EXPERIMENTAL EVALUATION

In this section we present experimental design and evaluation of Albanian question classification models. We describe the objectives and outcomes of each experiment and highlight performance variations between classical machine learning algorithms, with deep learning models. Classification performance is evaluated using the common metrics of accuracy, precision, recall and F-score.

Experiment 1: Question Classification using the classical machine learning algorithms.

The implementation of FastText for vector representation is shown to increase the classification performance in all the classical machine learning models used for classification [5]. In Table V we show the performance of Support Vector Machine, Random Forest, and Logistic Regression where FastText is used to generate word embeddings. Support Vector Machine proves to be the best performing algorithm among Random Forest and Logistic Regression, with scoring 75.7% for accuracy and 80.0% for F1-score which is not satisfactory if we compare it with state-of-the-art Question Classification models for languages other than Albanian. However, classical Machine Learning algorithms are not time-consuming and require less powerful processing units.

TABLE V. CLASSIFICATION PERFORMANCE IN THE PRE-PROCESSED ALBANIAN DATASET

Models (Albanian)	Accuracy	F1-score macro	Precision	Recall
Support Vector Machine (SVM)	**75.7%**	**80.0%**	**71.6%**	**73.9%**
Logistic Regression (LogReg)	74.0%	79.6%	64.9%	66.1%
Random Forest (RnFor)	70.8%	78.6%	62.5%	64.9%

Experiment 2: Deep learning classification models using the preprocessed Albanian dataset.

The main goal of this experiment is to show the effects of preprocessing on question classification task in Albanian. As discussed in Section IV, as a preprocessing step we chose the stop-word removal. The most frequent Albanian words were removed from the questions dataset. The same Deep Learning models were trained using the preprocessed Albanian dataset. The results are shown in Table VI. RoBERTa proves to be the best performing model compared to BiLSTM, BiLSTM with attention and Transformer.

TABLE VI. DL CLASSIFICATION PERFORMANCE IN THE PRE-PROCESSED ALBANIAN DATASET

Models (Albanian)	Accuracy	F1-score macro	Precision	Recall
BiLSTM	80.8%	78.5%	81.7%	76.6%
BiLSTM + Attention	81.5%	81.0%	86.2%	78.2%
Transformers	85.3%	84.0%	85.1%	83.0%
RoBERTa	**87.6%**	**86.5%**	**85.5%**	**88.2%**

Experiment 3: Deep learning classification models using the Albanian dataset without prior preprocessing.

Questions are sentences with a specific structure that contains a question word which is classified a stop-word. The motivation behind this experiment is to measure the effect of stop-words in question classification in Albanian. The same models as in Experiment 2 were trained and tested on the Albanian questions dataset without removing stop-words from the dataset. The evaluation metrics of the models are shown in Table VII. RoBERTa outperforms the classification performance of other Deep Learning architectures.

In Fig. 2 we analyze the results of Experiment 2 and Experiment 3. Except from the Transformer, that is not significantly affected by stop-word removal, the other models show meaningful performance improvement when the stop-words are preserved in the dataset. When the stop-words were not removed, the accuracy of RoBERTa increased by 3.5%.

TABLE VII. CLASSIFICATION PERFORMANCE WITHOUT PRE-PROCESSING ALBANIAN DATASET

Models (English)	Accuracy	F1-score macro	Precision	Recall
BiLSTM	82.5%	81.3%	85.2%	79.2%
BiLSTM + Attention	83.2%	82.1%	83.0%	81.4%
Transformers	84.4%	82.3%	81.3%	83.4%
RoBERTa	**91.1%**	**90.0%**	**90.0%**	**90.2%**

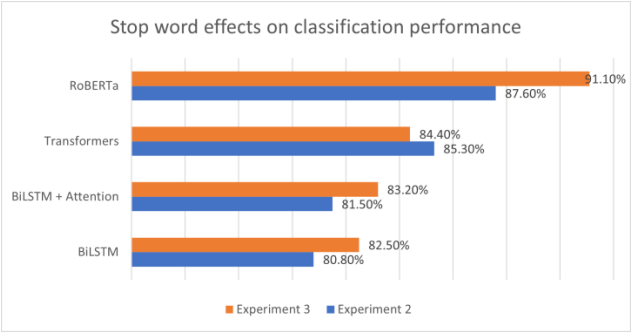

Fig. 2. Accuracy of DL models in experiments 2 and 3.

Experiment 4: Question Classification of deep learning models using the original TREC-6 dataset in English.

The main objective of this experiment is to determine a benchmark for the question classification performance. Albanian is an understudied language and the pretrained models are based on raw data. Furthermore, the automatic translation used to create the Albanian dataset, may negatively affect the performance of our model. English is a computationally rich and very well studied language. In order to assess if we have a decrease in performance due to low resources in Albanian language, we will define the state-of-the-art performance by training the models with the TREC dataset in English.

In Table VIII, we show the evaluation metrics calculated on our Deep Learning models trained with the original TREC-6 dataset. We show an increase in performance by adding attention to BiLSTM model. Furthermore, we confirm the superiority of Transformer in comparison to BiLSTM. The best performing classification is achieved using the RoBERTa pre-trained model with a benchmark from F-measure scoring 95.1%.

TABLE VIII. DL MODELS TRAINED WITH TREC-6 DATASET

Models (English)	Accuracy	F1-score macro	Precision	Recall
BiLSTM	81.8%	81.7%	80.9%	82.8%
BiLSTM + Attention	83.2%	83.2%	86.2%	81.3%
Transformers	86.0%	86.4%	88.1%	85.2%
RoBERTa	**94.0%**	**95.1%**	**95.0%**	**95.3%**

Experiment results show that Support Vector Machines model performs better compared to Random Forest and Logistic Regression, but its classification performance is still poor, with F1-score of 80%.

Among the deep learning models tested with the Albanian corpus dataset, RoBERTa performs better compared to BiLSTM and Transformer with an accuracy of 87.6%. We must state that all deep learning models performed better than SVM, Random Forest and Logistic Regression.

Stop word removal pre-processing task, applied on the Albanian questions dataset, shows a decrease in performance of the all the deep learning classifiers. Question classification with RoBERTa has an accuracy of 87.7% when trained with the dataset where the stop-words were previously removed, and the accuracy improved to 91.1% when trained with the dataset where the stop-words were not removed. We conclude that stop-words should not be removed from Albanian corpus prior to question classification.

The benchmark of 95.1% for the accuracy obtained by RoBERTa using the original TREC dataset in English shows that the performance of the question classification system in Albanian can be improved by working with the Albanian language from a computational linguistic perspective.

VII. CONCLUSION AND FUTURE WORK

Question Classification is a fundamental part of a Question Answering system. In this paper we have modeled the Question Classification problem in an Albanian language corpus using the state-of-the-art Deep Learning models and architectures. To the best our knowledge there is no previous research work addressing the problem of question classification in Albanian language using recent deep learning approaches. We have employed an Albanian questions dataset to train classification models based on classical machine learning algorithms and more recent deep learning approaches. The experiments, both with and without stop-words, demonstrate that the presence of stop-words significantly affects the accuracy of the classifier. Moreover, the comparison between algorithms indicates that deep learning algorithms outperform conventional machine learning approaches.

As future work, we intend to expand the size of the dataset and also increase its overall quality by adding linguistic

expertise. In addition, we plan to analyze the impact of the length of the questions on the classifier performance. Finally, it would be interesting to investigate language alignment approaches in order to exploit well-established machine translation algorithms in the process of classification.

REFERENCES

[1] Plepi, J., Kacupaj, E., Singh, K., Thakkar, H. and Lehmann, J., "Context Transformer with Stacked Pointer Networks for Conversational Question Answering over Knowledge Graphs," in European Semantic Web Conference, LNISA,volume 12731, 2021.

[2] Kacupaj, E, Plepi, J., Singh, K., Thakkar, H. and Lehmann, J., "Conversational Question Answering over Knowledge Graphs with Transformer and Graph Attention Networks," in 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 850–862, ACL, 2021.

[3] Alanazi, S. S., Elfadil, N., Jarajreh, M. and Algarni, S., "Question Answering Systems: A Systematic Literature Review," International Journal of Advanced Computer Science and Applications, vol. 12, no. 3, pp. 495-502, 2021.

[4] Sati, A. B. B., Ali, M. A. S. and Abdou, Sh. M., "Arabic Text Question Answering from an Answer Retrieval Point of View: a survey," International Journal of Advanced Computer Science and Applications, vol. 7, no. 7, pp. 478-484, 2016.

[5] Kote, N., Trandafili, E. and Plepi, Gj., "Question Classification for Albanian Language: An Annotated Corpus and Classification Models," in Advances on P2P, Parallel, Grid, Cloud and Internet Computing. 3PGCIC, Tirana, 2022.

[6] Trandafili, E., Mece, E., Kica, K. and Paci, H., "A Novel Question Answering System for Albanian Language.," in Proceedings of EIDWT 2018, Tirana, Albania, 2018.

[7] Kodra, L. and Kajo Meçe, E., "Question Answering Systems: A Review on Present Developments, Challenges and Trends," International Journal of Advanced Computer Science and Applications, vol. 8, no. 9, pp. 217-224, 2017.

[8] Cortes, E. G., Woloszyn, V., Binder, A., Himmelsbach, T., Barone, D. and Moller, S., "An Empirical Comparison of Question Classification Methods for Question Answering Systems," in Proceedings of the 12th LREC, Marseille, 2020.

[9] A. Mohasseb, M. Bader-El-Den and M. Cocea, "Domain Specific Grammar based Classification for Factoid Questions," in Proceedings of 5th International Conference on Web Information Systems and Technologies, Vienna, Austria, 2019.

[10] A. Davidescu, A. Heyl, S. Kazalski, I. Cramer and D. Klakow , "Classifying German Questions According to Ontology-Based Answer Types," in Advances in Data Analysis, Berlin, 2007.

[11] M. Pota, M. Esposito and G. De Pietro, "Convolutional Neural Networks for Question Classification in Italian Language," in The 16th International Conference on Intelligent Software Methodologies, Tools, and Techniques (SOMET_17), Japan, 2017.

[12] A.-L. Ligozat, "Question Classification Transfer," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 2013.

[13] M. A. Garcıa Cumbreras, L. A. Urena Lopez and F. Martınez Santiago, "BRUJA: Question Classification for Spanish.Using Machine Translation and an English Classifier," in EACL 2006 Workshop on Multilingual Question Answering - MLQA06, 2006.

[14] X. Li and D. Roth, "Learning Question Classifiers," in COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[15] D. Metzler and W. B. Croft, "Analysis of Statistical Question Classification for Fact-based Questions," Information Retrieval, vol. 8, no. 3, p. 481–504, 2005.

[16] Â. Costa, T. Luís, J. Ribeiro and A. Crist, "An English-Portuguese parallel corpus of questions," in Proceedings of LREC'12, Istanbul, Turkey, 2012.

[17] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo and M. de Rijke, "Creating the DISEQuA Corpus: A Test Set for Multilingual Question Answering," in Comparative Evaluation of Multilingual Information Access Systems, 2003.

[18] E. Hovy, U. Hermjakob and D. Ravichandran, "A Question/Answer Typology with Surface Text Patterns," in HLT '02: Proceedings of the second international conference on Human Language Technology Research, 2002.

[19] A. Aouichat, M. S. Hadj Ameur and A. Geussoum, "Arabic question classification using support vector machines and convolutional neural networks," in Natural Language Processing and Information Systems, 2018.

[20] S. Xu, G. Cheng and F. Kong, "Research on Question Classification for Automatic Question Answering," in 2016 International Conference on Asian Language Processing (IALP), 2016.

[21] T. Hao, W. Xie, Q. Wu, H. Weng and Y. Qu, "Leveraging question target word features through semantic relation expansion for answer type classification," Knowle dge-Base d Systems, vol. 133, pp. 43-52, 2017.

[22] Y. Liu, X. Yi, R. Chen, Z. Zhai and J. Gu, "Feature extraction based on information gain and sequential pattern for English question classification," The Institution of Engineering and Technology, vol. 12, no. 6, pp. 520-526, 2018.

[23] A. Mohasseb, M. Bader-El-Den and M. Cocea, "Classification of factoid questions intent using grammatical features," ICT Express, vol. 4, no. 4, pp. 239-242, 2018.

[24] M. Pota and M. Esposito, "Question Classification by Convolutional Neural Networks Embodying Subword Information," in International Joint Conference on Neural Networks (IJCNN), 2018.

[25] G. Di Gennaro, A. Buonanno, A. Di Girolamo, A. Ospedale and F. Palmieri, "Intent Classification in Question-Answering Using LSTM Architectures," in Progresses in Artificial Intelligence and Neural Systems. Smart Innovation, Systems and Technologies, vol 184., Singapore, Springer, 2020, pp. 115-124.

[26] X. Luo, H. Ding, M. Tang, P. Gandhi, Z. Zhang and Z. He, "Attention Mechanism with BERT for Content Annotation and Categorization of Pregnancy-Related Questions on a Community Q&A Site," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, 2020.

[27] C. Mallikarjuna and S. Sivanesan, "Question classification using limited labelled data," Information Processing & Management,, vol. 59, no. 6, 2022.

[28] S. Raschka, J. Patterson and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," Information 11, vol. 4, no. 193, 2020.

[29] J. M. Perkel, "Why Jupyter is data scientists' computational notebook of choice," Nature, vol. 563, no. 7729, pp. 145-146, 2018.

[30] W. McKinney, "Pandas: a foundational Python library for data analysis and statistics.," Python for high performance and scientific computing, vol. 14, no. 9, pp. 1-9, 2011.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapea, M. Brucher, M. Perrot and É. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825-2830, 2011.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steine, L. Fang, J. Bai and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems 32, 2019.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need.," in Advances in neural information processing systems 30, 2017.

[34] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT 2019, 2019.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[36] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[37] Y. HaCohen-Kerner, D. Miller and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," PLoS ONE, vol. 15, no. 5, 2020.

[38] V. Kocaman and D. Talby, "Spark NLP: Natural Language Understanding at Scale," Software Impacts, vol. 8, 2021.

[39] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Subword Information.," Transactions of the Association for Computational Linguistics, vol. 5, p. 135–146, 2017.

[40] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in In Proceedings of the IEEE International Joint Conference on Neural Networks(IJCNN), Montreal, QC, Canada, 2005.

[41] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in International Conference on Learning Representations (ICLR), San Diego, 2015.