

Natural language processing for Albanian: a state-of-the-art survey

Muhamet Kastrati, Marenglen Biba

Department of Computer Science, Faculty of Engineering and Architecture, University of New York Tirana, Tirana, Albania

Article Info

Article history:

Received Jul 27, 2021

Revised Jun 8, 2022

Accepted Jul 2, 2022

Keywords:

Albanian language

Deep learning

Emotion detection

Machine learning

Natural language processing

Sentiment analysis

ABSTRACT

Due to its wide applicability, natural language processing (NLP) has attracted significant research efforts to the machine learning and deep learning research community. Despite this, research works investigating NLP for the Albanian language are still limited. However, to the best of our knowledge, there is no literature review available, which presents a clear picture of what has been studied, argued, and established in the area. The main objective of this survey is to comprehensively review, analyze and discuss the state-of-the-art in NLP for the Albanian language. Here, we present an extensive study concerning the contribution of several authors that have contributed to the application of NLP to the Albanian language. Also, we present an overview of research carried out in the typical applications of NLP for the Albanian language. Finally, some future challenges and limitations of the area are discussed.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Muhamet Kastrati

Department of Computer Science, Faculty of Engineering and Architecture, University of New York Tirana
Kodra e Diellit, Tirana, Albania

Email: muhamet.kastrati@gmail.com

1. INTRODUCTION

Natural language processing (NLP) is an emerging interdisciplinary research area at the intersection of learning linguistics, computer science, and artificial intelligence (AI) mainly dealing with computational techniques to learn, understand, and produce human language content [1]. Over the last decades, the NLP community has been focused on topics such as machine translation, speech recognition, and speech synthesis. During the last 20 years, there has been an increased interest among the NLP research community and practitioners in a wide range of real-world applications including speech-to-speech translation engines, mining social media for information about health and finance, emotion and sentiment analysis concerning products, and/or services [1]. Historically, there have been known three main approaches used in NLP: rule-based, machine learning, and deep learning approach. The rule-based approach is the earliest type of all AI algorithms. Second, is the machine learning approach that covers several machine algorithms that have been used to solve several NLP tasks. Third, is the deep learning approach or deep neural networks (DNN), which have revolutionized the field of NLP [2]. Among the most successful NLP tasks tackled by the research community and practitioners includes named entity recognition, text classification, topic modeling, parts-of-speech tagging, question answering, chatbots, image captioning, fake news detection, text generation, sentiment and emotion analysis, speech-to-text, text-to-speech, and topic classification. Despite the advancement of research in NLP for other high-resource languages, such as English, French, Spanish, German, and Chinese, much less work has been done so far on other low-resource languages where Albanian belongs to.

More recently, there has been increased research interest among researchers to explore NLP for the Albanian language mainly dealing with tools and applications. Studies conducted concerning tools include

part-of-speech and morphological tagging [3]–[5], stemming [6], the lexicon of Albanian for NLP [7], and syntactic parsing [8]. Also, there are several studies related to the application of the NLP for the Albanian language such as named-entity recognition [9]–[12], sentiment analysis [13]–[16], emotion detection [17], [18], hate speech detection [19], [20], summarization techniques [21], [22], text classification [23]–[25] and question answering system [26].

To the best of our knowledge, this survey is the first one on this topic (NLP for the Albanian language). The literature used for this survey encompasses almost all research works published in conference proceedings and journals among other sources used in NLP for the Albanian language area. All papers reviewed in this survey are considered from two main perspectives, respectively, the technical approach used for learning and the NLP tasks that have been tackled.

The rest of the paper is structured as the following: section 2 presents a brief theoretical background on NLP and the Albanian language. Section 3 presents a comprehensive state-of-the-art review in NLP for the Albanian language from its beginning to the present day. Section 4 concludes the paper.

2. BACKGROUND

In order to facilitate understanding, we introduce a theoretical background on NLP and the Albanian language that are part of this survey. NLP is a subfield of computer science that employs computational techniques to learn, understand, and produce human language content [1]. Albanian is an Indo-European (IE) language, an independent branch of its own, which has distinctive features that range from morphological to lexical viewpoints.

2.1. Natural language processing

NLP is an emerging interdisciplinary research area at the intersection of learning linguistics, computer science, and artificial intelligence mainly dealing with the interaction between computers and human language. From their beginnings, the main goal of the NLP systems has been aiding human-human and human-machine communication. In the first case, human-human communication, one typical example is the case of machine translation (MT). In the second case, human-machine communication, some common examples include conversational agents. In general, both humans and machines benefit from human language content data that is nowadays available online [1]. Over the past two decades, there has been an increased interest by both the research community and practitioners as the NLP has been shown to perform quite well in several consumer products (for example, in speech-based natural user interfaces (NUI) such as Alexa, Cortana, Google Assistant, and Apple's Siri). As described by Hirschberg and Manning [1], the success of NLP is strongly related to these four key factors: i) advances in computation power, ii) a large amount of available natural language content, iii) advancements in machine learning and deep learning respectively, and iv) a better understanding of the human languages including grammatical structure and meaning.

2.2. Albanian language

Albanian, in its modern form, is the official language of Albania, Kosovo, and North Macedonia, where it has co-official status. Albanian is currently spoken by more than 7.5 million Albanians living in Albania, Kosovo, Montenegro, northwest Macedonia, some other Western European countries, and North America where Albanian people live and work. The Albanian language is considered an Indo-European (IE) language family, which derives from the old Illyrian language. Same as the Greek and Armenian language it is an independent branch of IE language, which has been mainly spoken in the Western Balkans. Same as other languages, also the Albanian language has distinctive features that range from morphological to lexical viewpoints.

3. STATE-OF-THE-ART REVIEW

In this section, we have presented a state-of-the-art review of studies conducted concerning the NLP tools and applications for the Albanian language. It starts with some early studies conducted about parts-of-speech and morphological tagging, annotated corpora, and then followed by a brief overview of the stemming, and parsing. Finally, a short overview of the studies conducted concerning sentiment analysis, emotion detection, hate speech detection, text classification, question answering systems, text classification, and named entity recognition for the Albanian language is given.

3.1. Part-of-speech and morphological tagging

Over the last years, there have been several attempts to build NLP tools for the Albanian language. In the following, we will briefly overview some of the most significant research work conducted in this area.

Almost all these existing tools are rule-based or dictionary-based and unfortunately are not available online for NLP purposes.

Trommer and Kallulli [3] presented a simple morphological tagger for the standard Albanian language. All that is intended here was just to introduce an initial component of an annotation tool in the context of the Albanian Corpus Initiative. Their proposed morphological tagger has been evaluated in a small corpus of 1,000 tokens (words) and the results obtained were very promising, with a Precision of 97% and Recall in the range of 92–95%.

Several studies [27], [28] presented and developed electronic dictionaries and transducers for the automatic processing of the Albanian language. The authors described some peculiarities of the Albanian language and then explained how FST and generally speaking NooJ's graphs enable to treat them. Their focus was to analyze the words inside a linear segment of text. They also studied the relationship between units of sense and units of form.

3.2. Annotated corpora

Arkhangelskij *et al.* [29] presented a large annotated corpus of Albanian created by the Saint-Petersburg linguists' team. The corpus consists of 16.6 M tokens collected from several sources including textbooks, news reports, official, religious, and scientific texts. Kadriu in [30] presented an unsupervised approach using a dictionary with around 32,000 words with the corresponding part-of-speech (POS) tags of the words and a set of regular expression rules to assign POS tags to a new text, using the natural language toolkit (NLTK) toolkit. The evaluation has been performed on a set of 30 random news articles, and the accuracy reached was in the range of 88 to 93%. Some other annotated corpora have been presented by the UniMorph project [31], which is a small annotated morphological corpus of Albanian inflected words extracted from Wiktionary.

Kabashi and Proisl in [5] presented the gold version corpus of their previous corpus introduced in Kabashi and Proisl [4]. Their corpus contains 2,020 sentences, with 31,584 tokens and was manually annotated by two native Albanian speakers. The obtained results reached an accuracy ranging from 86.96% to 95.10%.

3.3. Stemming and parsing

Karanikolas in [32] provided a naive-single-step (rudimentary) stemming algorithm for the Albanian language. Here the author also presented a list of 470 stop words and a corpus containing 5,000 words, which was used to generate the stems. The obtained results reached an accuracy of 80%. It is good to emphasize that the author and the evaluators engaged in this study were not native speakers. Sadiku and Biba [6] presented the first stemming algorithm for the Albanian language developed by a native speaker. Here authors built upon a rule-based JStem algorithm, which is based on word formation with affixes.

Misini *et al.* in [8] provided an appropriate method for syntactic parsing of the Albanian language. In this article, the authors begin by describing the prior work and pointing out several algorithms used for parsing. They then state that parsing is the most appropriate approach to identify the syntactic structure that is useful in determining the meaning of a sentence. Their algorithm is based on the idea of splitting the sentences into parts of speech and analyzing these sentences using the natural language's syntactic rules.

Collaku and Adali [33] introduced a new method of grouping verbs based on their inflection themes. Contrary to traditional classification methods, where verbs are classified based on the inflection themes they take, here verbs are classified into different verb groups. By doing this, the inflection process looks clearer and more regular, as the affix remains the only changeable part of the inflected verb. The most important outcome of this approach is that it makes the process of Albanian verbs simpler and easier.

3.4. Sentiment analysis and emotion detection

In this section, we will survey scientific literature concerning sentiment analysis and emotion detection tasks for the Albanian language. As described by authors in [34] in general, research about sentiment analysis has been done at three levels: document level, sentence level, and aspect level sentiment classification. Over the last few years, sentiment analysis is one of the most active research areas in NLP [35]. However, there are only a few researches dedicated to sentiment analysis (opinion mining) for the Albanian language presented by authors in [13]–[16] and a few others related to emotion detection in studies [17], [18].

Biba and Mane [13] presented the first approach for sentiment analysis in Albanian. They developed a machine learning model to classify text documents belonging to a negative or positive opinion regarding the given topic. To train their machine learning models they built a corpus of 400 documents containing political news consisting of five different topics. In this case, each topic was represented by 80 documents classified as positive or negative. Here authors made an empirical comparison between 6 different machine

learning algorithms, respectively, Bayesian logistic regression, logistic regression, support vector machine (SVM), voted perceptron, naive Bayes, and hyper pipes for classification of text documents, achieving accuracy between 86% and 92% depending on the topic. The authors conclude that, in general, to achieve higher accuracy in sentiment analysis, a larger corpus in the Albanian language is needed.

Kote *et al.* [14] presented a comprehensive experimental evaluation of machine learning algorithms applied for opinion mining in the Albanian language. Among the algorithms tested that performed better include logistic and multi-class classifier, hyper pipes, radial basis function (RBF) classifier, and RBF network with the classification accuracy ranging from 79% to 94%. Here authors trained the classification algorithms over a corpus of 500 news articles in Albanian consisting of 5 different topics. Here, each topic was represented with a balanced set of articles opinionated as positive or negative. The experimental results were interpreted concerning several evaluation criteria for each algorithm showing interesting features in the performance of each algorithm.

Kastrati *et al.* [15] presented the sentiment analysis of people's opinions expressed on Facebook with regard to the pandemic situation in the Albanian language. Here the authors developed a deep learning-based model to classify people's opinions as neutral, negative, or positive. In order to train their model, they created a new specific corpus containing 10,742 manually annotated Facebook comments in the Albanian language. The authors conclude that combining the bidirectional long short-term memory neural network (BiLSTM) with an attention mechanism outperformed the other approaches in their sentiment analysis task. The best results achieved by their proposed model given in terms of Precision, Recall, and F1-score were 72.31%, 72.25%, and 72.09%, respectively.

Vasili *et al.* [16] presented sentiment analysis on Twitter messages for the Albanian language. The authors compared the results among several methods and noted the challenges that arise when dealing with sentiment analysis for Albanian language. They studied the performance of sentiment classification techniques using three main approaches: traditional machine learning, lexicon-based and deep learning-based approach. Their experiments revealed that long short-term memory (LSTM) based recurrent neural network (RNN) with Glove as a feature extraction technique provides the best results with F-score=87.8%, followed by Logistic Regression.

Skënduli and Biba [17] presented an approach for analyzing users' emotions on microblogging texts and postings in the Albanian language. Here authors studied users' emotions at the sentence level by using deep learning methods. Their approach was based on the idea of classifying a text fragment into a set of pre-defined emotion categories (based on Ekman's model) and therefore aims at detecting the emotional state of the writer conveyed through the text. To perform their experiments, they built a new dataset that contained manually annotated Facebook posts belonging to some active Albanian politicians, which were classified using Ekman's model and finally separated into six smaller datasets. Then, they performed a set of experiments on these datasets to evaluate deep learning and classical machine learning models. Furthermore, in their analysis, they also adopted a domestic stemming tool for the Albanian language to preprocess the datasets, which showed a slight improvement in the classification accuracy. In general, the obtained results showed that deep learning outperformed the other classical machine learning models, that is, Naive Bayes (NB), instance-based learner (IBK), and support vector machines (SMO) with a given accuracy ranging from 70.2% to 91.2% for (correctly classified unstemmed instances), and 67.0% to 92.4% for (correctly classified stemmed instances).

3.5. Hate speech detection

Internet in general and online social media, in particular, have greatly facilitated and changed the way people communicate. As generally there is no censorship, online social media platforms sometimes are used for the dissemination of aggressive and hateful content. Recently, there has been an increased research interest among the academic community and practitioners to build effective automatic solutions for hate speech detection. However, there is very limited work to address this problem for the Albanian language.

Ajdari *et al.* [19] explored the idea of building automatic hate speech detection in the public Albanian language pages. As there was no previous hate speech corpus for the Albanian language, they collected data from Facebook pages in the Albanian language and built their hate speech corpus. The corpus comprised 4,886 comments, and two annotators were employed to categorize all these comments as hate or no hate. Then, to assess the model generalization, the authors trained a SVM classifier using a training set of 4,000 instances, and the rest of 886 instances were used as a testing set. The best-obtained results by their proposed model in terms of precision 61%, recall 57%, and F1-score 58%. Further, Raufi and Xhaferri [20] attempted in the direction of implementation of a lightweight machine learning classification model for hate speech detection in the Albanian language for mobile applications. Their initial testing and evaluations provided promising results concerning classifier accuracy in mobile environments where frequent and real-time training of the algorithm is required.

3.6. Question answering system

Recently, there has been an increasing trend toward the implementation of different systems and tools for question answering. However, to the best of our knowledge, there is only one research work that addressed this problem for the Albanian language. Trandafilí *et al.* [26] work is the first study published around question answering system for the Albanian language. The system was built based on the idea of extracting answers to factoid questions for a given text. Experimental results obtained from the proposed approach were promising and showed that this was an effective solution for single domain documents [26].

3.7. Summarization techniques

Trandafilí *et al.* [22] proposed a novel document summarization system designed specifically for the Albanian language. Here authors showed experimentally that the enrichment of the summarization system with language-dependent elements improves the systems' performance and the compression rate. Vasili *et al.* [21] presented a study of summarization techniques for the Albanian language. Here, the authors begin by pointing out a theoretical approach where the widely used summarization techniques are described. They then further continued by applying these techniques to the Albanian language, since the language is an important factor that might lead to different outcomes for each algorithm, due to its structure, its form, and its rules. They also provided a comprehensive overview of the text summarization techniques mainly used for high-resource languages and then conclude about the most appropriate summarizing techniques for the Albanian language. They also provided a formal way of verifying the correctness of their obtained results, by using the best criteria defined by experts in the field. They also published their dataset and practical approach implemented in their study.

3.8. Text classification

Trandafilí *et al.* in [25], evaluated the performance of some of the most important text classification algorithms over a corpus composed of Albanian texts. The authors used several NLP preprocessing steps to their data before feeding them as input to the learning models. To assess the model generalization, the authors trained several algorithms, namely, simple logistics (SL), naive Bayes (NB), k-nearest neighbor (K-NN), decision trees (DT), random forest (RF), support vector machine (SVM), and neural networks (NN). The obtained results showed that naive Bayes and SVMs achieved better results compared to the other algorithms used to classify Albanian corpus.

Kadriu and Abazi in [23] evaluated and compared several classification algorithms used for text classification. Their research was mainly focused on the classification of text extracted from Albanian news articles into a set of pre-defined categories, namely, latest news, economy, sport, showbiz, technology, culture, and world. Authors here applied several preprocessing steps to the text corpus before feeding them as input to the machine learning models, including stop words removal, and the creation of a separate file for each category, then these files were split into sentences, then, for each sentence, one of the predefined categories was assigned, and finally, a list of tuples sentence/category has been created. In this study, the authors trained several classifiers including multinomial, linear support vector classifier (SVC), Neighbour, Bernoulli, Centroid, stochastic gradient descent (SGD), Perceptron, Ridge, passive aggressive. The average accuracy for the above classifiers was in the range of 74% and 91% depending on the input set.

Later on, Kadriu *et al.* [24] presented an extended study concerning text classification for Albanian news articles, which was based on a bag of words model and word analogies. In this study, the authors focused on text classification using two approaches: i) the text classification treats words as independent components and ii) the text classification treats words based on their semantic and syntactic word similarities. The obtained results showed that the bag of words model does better than fastText for smaller datasets and fastText showed better performance when classifying multi-label text. The best results were achieved with a bag of words model, with an accuracy of 94%.

3.9. Named entity recognition

Skënduli and Biba [9] proposed the first named entity recognition (NER) model for the Albanian language. The authors here employed a maximum entropy approach based on the Apache OpenNLP tool. For evaluation, they created a manually annotated corpus containing text from historical and political domains. Authors experimentally demonstrated that models can be further improved if larger corpora will be available. The obtained results reached the values of precision, recall, and F-measure 85%, 70%, 76% (person corpus), 83%, 66%, 73% (location corpus) and 69%, 60%, 64% (organization corpus) respectively.

Kono and Hoxha in [10] presented their NER approach for documents written in the Albanian language. They explored the use of conditional random fields (CRFs) for this purpose. They created a manually annotated corpus based on the Albanian news documents published in 2015 and 2016. The obtained results reached the values of precision, recall and F-score are 83.2%, 60.1%, and 69.7% respectively.

Hoxha and Baxhaku [11] reported on the first automatically-generated NE annotated corpus for the Albanian language. In this study, the authors have also used news articles from Albanian news media as a document source, which were automatically tagged using a custom generated gazetteer from Albanian Wikipedia.

Trandafili *et al.* in [12] proposed their NER approach for the Albanian language corpus based on deep learning models. In this study, the authors built a deep neural network using long short-term memory (LSTM) cells as hidden layers and CRF as the output, using both word and character tagging. For evaluation, they created their own manually annotated corpus. The results showed that the NER performance can be further improved by using a larger annotated corpus to train the model. Table 1 presents a short summary of the studies' characteristics on sentiment and emotion analysis, hate speech detection, question answering system, documents summarization, text classification, and named entity recognition for the Albanian language.

Table 1. A summary of the studies on the application of NLP for the Albanian language

Study	Application	Approach	Performance	Corpus
Biba and Mane [13]	Sentiment analysis	Machine learning	Accuracy: 86% to 92%	400 documents containing political news covering 5 different topics
Kote <i>et al.</i> [14]	Sentiment analysis	Machine learning	Accuracy: 79% to 94%	500 news articles covering 5 different topics
Kastrati <i>et al.</i> [15]	Sentiment analysis	Deep learning (BiLSTM + attention mechanism)	Precision: 72.31%, Recall: 72.25%, F1-score: 72.09%.	10,742 comments collected from the National Institute of Public Health of Kosovo (NIPHK)'s Facebook page
Vasili <i>et al.</i> [16]	Sentiment Analysis	Deep learning LSTM-RNN with Glove	Best results given in terms of F1 score F1: 87.8%	"Low quality annotators which should be eliminated from further considerations"
Skënduli-Prifti and Biba [17], [18]	Emotion analysis	Deep learning (CNN)	Accuracy: 70.2% to 91.2% (unstemmed) 67.0% to 92.4% for (stemmed)	6,358 Facebook posts belonging to Albanian politicians
Ajdari <i>et al.</i> [19]	Hate speech detection	Latent semantic analysis	Precision: 61%, Recall: 57%, F1-score: 58%.	4,886 Facebook posts
Raufi and Xhaferri [20]	Hate speech detection	Machine learning	Accuracy: 50% to 95%	421 Facebook posts, 648 words
Trandafili <i>et al.</i> [26]	Question answering	Machine learning	-	the first attempt at a Q&A system for Albanian
Trandafili <i>et al.</i> [22]	Document summarization	Machine learning	-	-
Trandafili <i>et al.</i> [25]	Text classification	Machine learning	-	-
Kadriu and Abazi [23]	Text classification	Machine learning	Avg accuracy: 74% and 91%	Albanian language news articles
Kadriu <i>et al.</i> [24]	Text classification	Machine learning	Accuracy: 94%	Albanian language news articles
Skenduli-Prifti and Biba [9]	Named entity recognition	Maximum entropy Markov model	Precision: 69% to 85% Recall: 60% to 70%	Tagged corpus of around 3,000 sentences and 87,900 words.
Kono and Hoxha [10]	Named entity recognition	CRFs	Precision: 83.26% Recall: 60.14% F-Score: 69.66%	50,000 words, from news articles
Hoxha and Baxhaku [11]	Named entity recognition	Maximum entropy model	Precision: 79.51% Recall: 40.55% F-Score: 52.99%	news articles from Albanian news media as a document source
Trandafili <i>et al.</i> [12]	Named entity recognition	Deep learning (LSTM and CRF as output)	Precision: 78.5% Recall: 72%	-





4. CONCLUSION

A major limitation of NLP today is the fact that most of the research work and application in industry is done only for the high-resource languages and there is less work for other low-resource languages including the Albanian language. This paper introduces and summarizes the most recent trends, methods, applications, and gaps, in the conducted research in the area of NLP for the Albanian language. Firstly, we described the theoretical basis of NLP, Albanian language, application of NLP for the Albanian language, and technical approaches used to tackle these tasks. Then we provided a comprehensive overview and research progress reached in NLP for the Albanian language. Finally, we discuss some challenges and open problems in NLP for the Albanian language.





REFERENCES

- [1] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015, doi: 10.1126/science.aaa8685.
- [2] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *arXiv:1702.01923*, 2017.
- [3] J. Trommer and D. Kallulli, "A morphological tagger for standard Albanian," in *Proceedings of LREC*, 2004, pp. 1–8.
- [4] B. Kabashi and T. Proisl, "A proosal for a part-of-speech tagset for the Albanian language," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4305–4310.
- [5] B. Kabashi and T. Proisl, "Albanian part-of-speech tagging: Gold standard and evaluation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 2593–2599.
- [6] J. Sadiku and M. Biba, "Automatic stemming of Albanian through a rule-based approach," *Journal of International Research Publications: Language, Individuals and Society*, vol. 6, 2012.
- [7] B. Kabashi, "A lexicon of albanian for natural language processing," *The XVIII EURALEX International Congress*, pp. 855–862, 2019.
- [8] A. Misini, E. Canhasi, and S. Krrabaj, "Albanian syntactic parsing," in *ICT Innovations 2020*, 2020, pp. 1–16.
- [9] M. P. Skenduli and M. Biba, "A named entity recognition approach for Albanian," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Aug. 2013, pp. 1532–1537, doi: 10.1109/ICACCI.2013.6637407.
- [10] G. Kono and K. Hoxha, "Named entity recognition in albanian based on crfs approach," in *Proceedings of the 2nd International Conference on Recent Trends and Applications in Computer Science and Information Technology, (RTA-CSIT)*, 2016, pp. 47–52.
- [11] K. Hoxha and A. Baxhaku, "An automatically generated annotated corpus for Albanian named entity recognition," *Cybernetics and Information Technologies*, vol. 18, no. 1, pp. 95–108, Mar. 2018, doi: 10.2478/cait-2018-0009.
- [12] E. Trandafil, E. K. Meçe, and E. Duka, "A named entity recognition approach for Albanian using deep learning," in *Complex Pattern Mining*, 2020, pp. 85–101.
- [13] M. Biba and M. Mane, "Sentiment analysis through machine learning: An experimental evaluation for Albanian," in *Recent Advances in Intelligent Informatics*, 2014, pp. 195–203.
- [14] N. Kote, M. Biba, and E. Trandafil, "A thorough experimental evaluation of algorithms for opinion mining in Albanian," in *Advances in Internet, Data & Web Technologies*, 2018, pp. 525–536.
- [15] Z. Kastrati, L. Ahmedi, A. Kurti, F. Kadriu, D. Murtezaj, and F. Gashi, "A deep learning sentiment analyser for social media comments in low-resource languages," *Electronics*, vol. 10, no. 10, May 2021, doi: 10.3390/electronics10101133.
- [16] R. Vasili, E. Xhina, I. Ninka, and D. Terpo, "Sentiment analysis on social media for Albanian language," *OALib*, vol. 08, no. 06, pp. 1–31, 2021, doi: 10.4236/oalib.1107514.
- [17] M. P. Skenduli, M. Biba, C. Loglisci, M. Ceci, and D. Malerba, "User-emotion detection through sentence-based classification using deep learning: A case-study with microblogs in Albanian," in *Foundations of Intelligent Systems*, 2018, pp. 258–267.
- [18] M. P. Skenduli and M. Biba, "Classification and clustering of emotive microblogs in Albanian: two user-oriented tasks," in *Complex Pattern Mining*, 2020, pp. 153–171.
- [19] J. Ajdari, F. Ismaili, B. Raufi, and X. Zenuni, "Automatic hate speech detection in online contents using latent semantic analysis," *Pressacademia*, vol. 5, no. 1, pp. 368–371, Jun. 2017, doi: 10.17261/Pressacademia.2017.612.
- [20] B. Raufi and I. Xhaferri, "Application of machine learning techniques for hate speech detection in mobile applications," in *2018 International Conference on Information Technologies (InfoTech)*, Sep. 2018, pp. 1–4, doi: 10.1109/InfoTech.2018.8510738.
- [21] R. Vasili, E. Xhina, I. Ninka, and T. Souliotis, "A study of summarization techniques in Albanian language," *Knowledge International Journal*, vol. 28, no. 7, pp. 2251–2257, Dec. 2018, doi: 10.35120/kij28072251R.
- [22] E. Trandafil, H. Paci, and E. Karaj, "A novel document summarization system for Albanian language," in *Proceedings of the 20th International Conference on Computer Systems and Technologies*, Jun. 2019, pp. 273–277, doi: 10.1145/3345252.3345275.
- [23] A. Kadriu and L. Abazi, "A comparison of algorithms for text classification of Albanian news articles," *Entrenova-Enterprise Research Innovation Conference*, vol. 3, no. 1, pp. 62–68, 2017.
- [24] A. Kadriu, L. Abazi, and H. Abazi, "Albanian text classification: Bag of words model and word analogies," *Business Systems Research Journal*, vol. 10, no. 1, pp. 74–87, Apr. 2019, doi: 10.2478/bsrj-2019-0006.
- [25] E. Trandafil, N. Kote, and M. Biba, "Performance evaluation of text categorization algorithms using an Albanian corpus," in *Advances in Internet, Data & Web Technologies*, 2018, pp. 537–547.
- [26] E. Trandafil, E. K. Meçe, K. Kica, and H. Paci, "A novel question answering system for Albanian language," in *Advances in Internet, Data & Web Technologies*, 2018, pp. 514–524.
- [27] O. Piton, K. Lagji, and R. Përmasa, "Electronic dictionaries and transducers for automatic processing of the Albanian language," in *Natural Language Processing and Information Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 407–413, doi: 10.1007/978-3-540-73351-5_38.
- [28] O. Piton and K. Lagji, "Morphological study of Albanian words, and processing with NooJ," *arXiv preprint arXiv:1002.0485*, Feb. 2010.
- [29] T. Arkhangelskij, M. Daniel, M. Morozova, and A. Rusakov, "Albanian and the languages of the Balkans," (in Albanian), in *Konferencë shkencore - Scientific Conference*, 2012, pp. 635–642.
- [30] A. Kadriu, "NLTK tagger for albanian using iterative approach," in *Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces*, 2013, pp. 283–288, doi: 10.2498/iti.2013.0565.
- [31] C. Kirov et al., "UniMorph 2.0: Universal Morphology," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 1–6.
- [32] N. N. Karanikolas, "Bootstrapping the Albanian information retrieval," in *2009 Fourth Balkan Conference in Informatics*, 2009, pp. 231–235, doi: 10.1109/BCI.2009.16.
- [33] I. Collaku and E. Adal, "Morphological parsing of albanian language: a different approach to albanian verbs," in *International Conference on Computer Science and Communication Engineering*, 2015, pp. 87–91, doi: 10.33107/ubt-ic.2015.94.
- [34] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/widm.1253.
- [35] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

BIOGRAPHIES OF AUTHORS

Muhamet Kastrati     received the B.Eng. degree in computer engineering from the University of Prishtina, Kosovo, in 2007 and an M.Sc. degree in computer engineering from the University of Prishtina, Kosovo, in 2014. Currently, he is a Ph.D. candidate at the department of computer science, University of New York Tirana, Albania. His research interests include optimization algorithms, statistical relational learning, machine learning, deep learning, natural language processing, and social network analysis. He can be contacted at email: muhamet.kastrati@gmail.com.



Marenglen Biba     received the Laurea Degree (5-year) Cumlaude degree in computer science from the University of Bari, Italy, in 2004, and the Ph.D. degree in computer sciences, in 2009. He has been a professor of computer sciences at the University of New York Tirana, since 2009. He is currently the Dean of the Faculty of Engineering and Architecture, University of New York Tirana in Albania. He has authored or co-authored several papers published in international journals and conferences and has served as a reviewer for many reputed journals. His research interests include artificial intelligence, machine learning, pattern recognition, data mining, computational biology, document image understanding, information extraction, and social network analysis. He can be contacted at email: marenglenbiba@unyt.edu.al.