Annals of Emerging Technologies in Computing (AETiC)
Vol. 5, No. 3, 2021

Research Article

Building Dictionaries for Low Resource Languages: Challenges of Unsupervised Learning

Diellza Nagavci Mati[1,*], Mentor Hamiti[1], Arsim Susuri[2], Besnik Selimi[1] and Jaumin Ajdari[1]

[1]South East European University, Tetovo, North Macedonia
dn16574@seeu.edu.mk; m.hamiti@seeu.edu.mk; b.selimi@seeu.edu.mk; j.ajdari@seeu.edu.mk
[2]University of Prizren, Prizren, Kosovo
arsim.susuri@gmail.com
*Correspondence: dn16574@seeu.edu.mk

Received: 24th February 2021; Accepted: 17th March 2021; Published: 1st July 2021

Abstract: The development of natural language processing resources for Albanian has grown steadily in recent years. This paper presents research conducted on unsupervised learning-the challenges associated with building a dictionary for the Albanian language and creating part-of-speech tagging models. The majority of languages have their own dictionary, but languages with low resources suffer from a lack of resources. It facilitates the sharing of information and services for users and whole communities through natural language processing. The experimentation corpora for the Albanian language includes 250K sentences from different disciplines, with a proposal for a part-of-speech tagging tag set that can adequately represent the underlying linguistic phenomena. Contributing to the development of Albanian is the purpose of this paper. The results of experiments with the Albanian language corpus revealed that its use of articles and pronouns resembles that of more high-resource languages. According to this study, the total expected frequency as a means for correctly tagging words has been proven effective for populating the Albanian language dictionary.

Keywords: ***Albanian language; corpora; dictionaries; natural language processing; part-of-speech tagging***

1. Introduction

Recent years have witnessed an increase in the interest regarding low-resource languages, and their need to be improved. Albanian is an Indo-European language spoken by around 7 million native speakers in Albania and Kosovo around the Balkans. The Albanian language is one of the most diverse and interesting languages to study due to its complex grammar and inflection paradigm; this makes morphological tagging extremely challenging [1]. In light of this, the Albanian language is one of the low-resources languages which has seen a gradual improvement year by year. One of the approaches to the Albanian language that can be evaluated is to use unsupervised learning methods that will learn from raw text. Unsupervised learning is the key to advancing the machine learning methods and unlocking access to almost unlimited amounts of data that can be used as training resources. Also, it involves training a model without pre-tagging or annotating [2].

The primary task of part-of-speech tagging is to take a text as an input and produce an output text where every word is marked with a mark corresponding to a grammatical category such as nouns, verbs, adjectives, adverbs, etc. This marking depends on the word's meaning and adverbs. Grammatical categories contain words that have the same grammatical properties.

The Albanian language has grammatical categories such as nouns, verbs, adjectives, numerals, genders and determinants, person-number indexing, tenses, active or passive voice. In grammar, a part of speech (also a word class, a lexical class, or a lexical category) is a linguistic category of words, which is generally defined by the syntactic or morphological action of the lexical item in question [11]. Tagging the

Diellza Nagavci Mati, Mentor Hamiti, Arsim Susuri, Besnik Selimi and Jaumin Ajdari, "Building Dictionaries for Low Resource Languages: Challenges of Unsupervised Learning", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 52-58, Vol. 5, No. 3, 1st July 2021, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.03.005, Available: http://aetic.theiaer.org/archive/v5/v5n3/p5.html.

Albanian language is especially challenging since it has extremely rich inflection paradigms and has 100 different forms: inflection patterns for levees of the same syntactic category. Research and comparative studies of the Albanian language are rarely conducted in Natural Language Processing. A small annotated morphological corpus of Albanian-inflected words extracted from Wiktionary with the (Universal Morphology) project was presented by Kirova *et al.* [14]. Annotations are done at the word level without regard to context and following the Universal Morphology schema, which associates each inflected word with its lemma and a set of morphological tags. There are 33,483-word forms for 589 lemmas in the corpus. Using the corpus, morphological analysis models have been trained and tested. Despite this, since the corpus contains individual words without sentence context, the data cannot be directly used to train a part-of-speech or morphological tagger.

In part-of-speech tagging, Kabashi *et al.* [15] proposed a set of part-of-speech tags after noticing that there was no set of moderate size. They built on their own previous work to achieve around 70% accuracy. Also, recently, Kote *et al.* [11] presented a corpus of 118,000 tokens tagged with part-of-speech and morphological features in Albanian. Furthermore, the team trained a neural morphological tagger and lemmatize that achieved good results, the best of which was 92.74% in POS tagging. In addition to using Universal Dependency guidelines to annotate the corpus, it went through a manual review process. In this study, 73 sources were tested at 77 MB of total capacity. In addition to the source frequency, 631,008 words have been tokenized and analyzed individually for frequency and different average frequency (△M). The Albanian corpus contains around 250,152 tokens, making it the largest so far. This paper examines the challenges of unsupervised learning of morphology for low resource languages in the Albanian language.

2. Natural Language Processing

Natural language processing's primary purpose is to understand and produce natural language at several levels: syntax, semantics, pragmatics, and dialogue. Among these levels are syntax, a language's structure, semantics, and pragmatics. A few steps are involved in Natural Language Processing, including tokenization, normalization, stemming, part-of-speech tagging, etc. [8]. Additionally, NLP includes different approaches and grammatical rules, such as inflection, derivation, tenses, semantic analyses, lexicon, morphemes, and corpora. In the domain-based corpora for the Albanian language, all these approaches and rules were applied.

2.1. Tokenization

As part of Natural Language Processing, there are several steps, among which is tokenization. The purpose of this step is to divide long strings into smaller chunks or tokens [5]. Massive chunks of text can be tokenized into sentences, so they can be tokenized into words. Furthermore, the processing is generally completed after a piece of text has been appropriately tokenized. Because the Albanian alphabet includes letters such as 'ç' and 'ë' in the text files, one challenge during tokenization is that UTF-8 is used to encode the text files. Otherwise, the files would not be appropriately tokenized.

2.2. Normalization

The next step of Natural Language Processing is normalization, which describes the process of pulling all text to a level playing field: converting all text to the same parameters such as uppercase or lowercase, removing punctuation. Some examples would be the roman numerals such as 'XX', 'V' etc.; hyphened words such as 'shoqërore-ekonomike' - Eng. 'socio-economic'; 'juridiko-civile' – Eng. 'civil-legal' etc.; words with apostrophes such as the word 'ç'është' abbreviated from 'çfarë është'-Eng. (what is), 't'i' shortened from 'të i' - Eng. (to) etc.; numbers expanding contractions, converting numbers to their word equivalents, and so on [4]. Normalization puts all words on a similar footing and allows processing to proceed uniformly.

2.3. Stemming

In Natural Language Processing, stemming is the process of removing affixes of suffixes, prefixes, infixes, and circumfixes from a word or phrase so that it may be processed further [2]. Examples of steaming in Albanian are shown in figure 1.

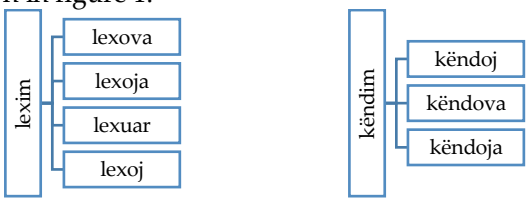

Figure 1. Stemming examples in the Albanian language

2.4. Corpus

Collection of the text refers to a corpus in Natural Language Processing. Such collections may be formed of a single language of texts or span many languages; there are numerous reasons for which multilingual corpora may be helpful. There have been several attempts to build a corpus for NLP tasks in the Albanian language. Still, they have significant drawbacks due to small corpus sizes, different formats fitted for individual tasks, and, moreover, they are not publicly available. The paper documents creating an Albanian corpus that includes 250k sentences from texts of various fields in electronic sources.

2.5. Part-of-speech (POS) Tagging

Part-of-speech tagging is a fundamental step in Natural Language Processing. The most popular part-of-speech tagging would be identifying words as nouns, verbs, adjectives, etc. The Albanian language has some properties that pose difficulties in creating a part-of-speech tag set [6]. A challenge faced in building a dictionary for low resource language (in this case for Albanian language) is that a part-of-speech tag set that can adequately represent the underlying linguistic phenomena is difficult to build. This difficulty is due to linguistic differences and a partial lack of standardization in using new/foreign words – especially in unedited electronic sources. B. Kabashi *et al.* has presented a corpus of approximately 2000 sentences by manually annotating them [3]. Some of those sentences have been selected randomly from texts of different genres. The remaining sentences have been selected manually to allow for a wider variety of linguistic phenomena in the sample corpus. The corpus presented in this research paper includes 250k sentences, which is the largest corpora proposed for the Albanian language. A tagger for the Albanian language is yet to be completed, but we report work to date.

Nouns have different morphological categories such as nominative, accusative, dative, genitive, and ablative: the inflection suffices for the dative, genitive, and ablative forms are identical. The distinction between them can only be made from context. On the other hand, as it pertains to indefiniteness (indefinite and definite): Indefinite forms like (një) vaja-Eng.(one/a) girl can be distinguished from definite forms like vajza - Eng. (the) girl [19].

The Albanian language nouns are determined by gender, such as feminine, masculine, and neutral- and numbers such as singular and plural. However, the majority of Albanian nouns change gender in the plural form. The words aktrim/aktrimi, for example, are masculine in the singular and feminine in the plural (aktrime/aktrimet). Hence they are known as heterogeneous nouns [7]. This phenomenon makes examining confluence difficult without a morphological component. Table 1 shows the proposed noun tags for the Albanian language.

Table 1. Albanian language proposed noun tags

Number	Name	Tag	Example in Albanian Language
1	Noun	No	marte
2	Noun word prep. art	NArt	marte; (të) rinjtë
3	Heterogeneous Noun	HgsNo .	aktrim, -i sg. m. vs. -e, -et pl. f.
4	Name	Nm	Tetova; Gea; Hana;

As an example, consider the sentence 'Të rinjtë janë nga Tetova.', Eng. 'The young people are from Tetova.', which is analyzed as Të\Art rinjtë\NArt janë\V nga\Prep Tetova\Nm.\Punct.

In the Albanian language, the same orphological categories are also applicable with adjectives. The meaning of adjectives in the Albanian language is to describe a person or thing in the sentence, which should also correspond to the gender and number of nouns. For instance, let us take a masculine and feminine example: "Ky është Andi, vëllau im."-Eng. "This is Andi, my brother."; whereas for feminine, it would be: "Kjo është Gea, motra ime."-Eng. "This is Gea, my sister.". From this, it becomes apparent that gender-wise there is a difference between 'ky' for masculine singular and 'kjo' for feminine singular – which in English is represented by 'this' in both genders [18].

Additionally, there is a difference in the personal pronoun between 'im', in masculine singular, additionally and 'ime' in feminine singular – which in English is represented by gender-neutral 'my'. The five proposed tags that can describe adjectives in the Albanian language are separate tags for adjectives that occur before nouns, adjectives with preposed articles, and for noninflectional-adjectives [20]. Table 2 below shows.

Table 2. Albanian language proposed adjective tags

Number	Name	Tag	Example in Albanian Language
1	Adjective	Adje	[vajza] mençur
2	Proposed adjective	PPAdje	mençuri [vajzë]
3	Adjective word article	AdjePpArt	[vajzë] e mirë.
4	PPAdj. word article	PPAdjePPArt	e mira [vajzë]
5	Noninflected adjective	AdjNIfe	blu/neto

Adjectives in the Albanian language have three forms: positive, comparative, and superlative [11]. Escalation is realized as a combination of the base word with the comparative article 'më', e.g. (1) positive: e bukur - Eng. 'beautiful', (2) comparative: 'më e bukur'-Eng. 'more beautiful' and (3) superlative: 'më e bukura' Eng. 'most beautiful'.

Numerals in the Albanian language are unclassified into cardinal and ordinary numbers. Ordinary numbers have the same properties as adjectives, except escalation, and are always preceded by an article.

Table 3. Albanian language proposed numeral tags

Number	Name	Tag	Example in Albanian Language
1	Cardinal number	NumCa	tre [fitore]
2	Ordinal number	NumOr	[fitorja] e tretë

As an example, consider the sentence: 'Kjo ishte fitorja e saj e tretë brenda një muaji.' - Eng. 'This was her third victory within one month.', which is tagged as Sot\Adv ishte\V fitorja\N e\Art tretë\NumO e\Art saj\PossPr brenda\Prep një\NumC muaji\No.\Punc.

In the Albanian language, pronouns are classified into subtypes according to their specificity. Furthermore, the relative pronoun is different from an interrogative pronoun or a personal pronoun [10]. Each distinguished type of pronoun has its own tag, as shown in Table 4 below.

Like nouns or adjectives, some pronouns can be preceded by a proposed article.

The interrogative pronoun, for example, 'cili' – Eng. 'who' can turn into 'which' when we consider relative the pronoun 'i cili', Eng. 'which'. This begs consideration that article and pronoun need to be treated together.

Table 4. Albanian Language proposed pronoun tags

Number	Name	Tag	Example in Albanian Language
1	Personal pronoun	PersPr	ti
2	Demonstrative pronoun	DemP	ky/kjo/këta
3	DemonstrativePron w. art	DemPPPArt	i tillë
4	Possesive pronun	PossPr	im
5	PossP w. prep. Art.	PossPPPArt	i tij/të vetën
6	Interrogative pron.	IntPr	kush
7	IntP w. art.	IntPPPrArt	i kujt/i cilit
8	Relative pronoun	RelPr	që
9	RelP w. art.	RelPPPrArt	i cili
10	Indefinite pron.	IndefPr	dikush
11	Reflexive pron.	ReflPr	(më)vete

3. Evaluation of Experiments

Using the Natural Language Toolkit for implementation of the corpus in Albanian, we conducted our experiments. NLTK is a group of open-source program modules for linguistic purposes. Natural language processing is covered by NLTK in both symbolic and statistical ways and is interfaced to annotated corpora. NLTK is the suit of Python modules that provide many Natural Language Processing data types, processing task, corpus simples, and problem sets [16-17]. In choosing Python as our object-oriented language, we wanted to make sure that data and code could be encapsulated and reused easily. A corpus was built by collecting all the texts in the Albanian language related to various fields, such as computer science, medicine, economy, politics, and tourism. Building a dictionary for the Albanian language is challenging, as has been discussed so far. The corpus has 631,008 words, of which 250,152 unique words were identified after removing duplicates and characters such as hyphens, apostrophes, numbers, and roman numerals. Table 5 shows some of the words derived from all sources used to calculate the expected frequency and average differences between them. An analysis of the highest source frequency, expected frequency and average differences from the total number of appearances resulted in the word 'të' having the highest expected frequency of 0.08033, an average difference of 0.013365, and source frequency of 73, and 588394 total appearances: continuing with the word 'e', 'në', 'vendet', 'thënë', 'shprehjes' etc.

Table 5. Average Difference for all sources

Word	Total (Expected Frequency)	Average Difference ΔM	Total number of appearances
të	0.08033	0.013365	588394
e	0.05662	0.007926	721846
në	0.03330	0.005964	249365
……			
vendet	0.00033	0.000382	4850
thënë	0.00033	0.000319	4841
shprehjes	0.00033	0.000622	4828
……			
tërkuzë	0.000001	0.000003	11
tinëzar	0.000001	0.000003	11
valëzim	0.000001	0.000003	11

A comparison between the obtained results of different languages, such as Albanian, English, and French, has also been conducted. Approximately 229,000 words and 11,150 tokens are included in the corpus of the English language, while the French language has 70,000 words and 9,150 tokens. In Table 6 below, we have summarized the most frequent English and French tokens to be compared with the Albanian language. The most frequent token among the Albanian language corpora is 'të' with 2.92%. The most frequent token among the English language corpora is 'the' with 3.24% in English. In the French language, the word 'de' is the most frequently used word with a frequency of 4.07%.

Table 6. Comparison of most frequent tokens in different languages.

Albanian Language		English Language		French Language	
Token	Source Frequency %	Token	Source Frequency %	Token	Source Frequency %
të	2.92%	the	3.24%	de	4.07%
e	2.92%	to	3.12%	la	2.95%
në	2.92%	of	2.65%	le	2.31%
thënë	1.83%	a	2.30%	et	2.12%
vendet	1.64%	I	2.23%	des	1.95%

4. Conclusion

This paper discusses the significance of the research presented in this corpus, which puts the Albanian language in a different linguistic context. A morphological tagger and 250,152 tokens are presented for the Albanian language. If the Albanian language is incorporated into the European

languages group, and more public documents are translated into this language, there will be plenty of scope for improving this corpus. We expect further improvement in tagger accuracy with more sources since we intend to train relatively good tagger models using a larger corpus to allow fully automatic tagging.

References

[1] Piotr Kłosowski, "Deep Learning for Natural Language Processing and Language Modelling", in *Proceedings of Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 19-21 September 2018, Poznan, Poland, Online ISBN: 978-8-3620-6533-2, E-ISBN: 978-83-62065-31-8, DOI: 10.23919/SPA.2018.8563389, pp. 22-32, Published by IEEE, Available: https://ieeexplore.ieee.org/document/8563389.

[2] Matteo Pagliardini, Prakhar Gupta and Martin Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features", in *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics,* June 2018, DOI: 10.18653/v1/N18-1049, pp. 528-540, Published by Association for Computational Linguistics (ACL), Available: https://www.aclweb.org/anthology/N18-1049/.

[3] Diellza Nagavci Mati, Mentor Hamiti, Besnik Selimi and Jaumin Ajdari, "Review of Natural Language Processing tasks in Albanian Language", in *Proceedings of the 3rd International Scientific Conference on Business and Economics,* 13-15 June 2019, Tetovo, North Macedonia, Online ISBN: 978-608-248-031-2, pp. 317-323, Available: https://www.researchgate.net/publication/348881531_Review_of_Natural_Language_Processing_tasks_in_Albanian_Language.

[4] Diellza Nagavci Mati, Mentor Hamiti, Jaumin Ajdari and Besnik Selimi, "A Systematic Mapping Study of Language Features Identification from Large Text Collection", in *Proceedings of the 8th Mediterranean Conference on Embedded Computing (MECO)*, 10-14 June 2019, Budva, Montenegro, Online ISBN: :978-1-7281-1740-9, E-ISBN: 978-1-7281-1739-3, DOI: 10.1109/MECO.2019.8760042, pp. 259-263, Published by IEEE, Available: https://ieeexplore.ieee.org/document/8760042.

[5] Besim Kabashi and Proisl Thomas, "Albanian Part-of-Speech Tagging: Gold Standard and Evaluation", in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC),* May 2018, Miyazaki, Japan, Online ISBN 979-10-95546-00-9, pp. 2593-2599, Published by European Language Resources Association (ELRA), Available: https://www.aclweb.org/anthology/L18-1412.

[6] Atta Ur Rahman, Khairullah Khan, Wahab Khan, Aurangzeb Khan and Bibi Saqia, "Unsupervised Machine Learning based Documents Clustering in Urdu", in *EAI Endorsed Transactions on Scalable Information Systems*, Print ISSN: 2032-9407, pp. 1-14, Volume 5, Issue. 19, June 2018, Published by European Alliance for Innovation, DOI: 10.4108/eai.19-12-2018.156081, Available: http://dx.doi.org/10.4108/eai.19-12-2018.156081.

[7] Arbana Kadriu, "NLTK Tagger for Albanian using Iterative Approach", in *Proceedings of the 35th International Conference on Information Technology Interfaces (ITI),* June 2013, Cavtat, Croatia, Online ISSN: 1334-2762, DOI: 10.2498/iti.2013.0565, pp. 283-288, Published by IEEE, , Available: https://ieeexplore.ieee.org/document/6649039.

[8] Emdad Khan, "Machine Learning Algorithms for Natural Language Semantics and Cognitive Computing", in *Proceedings of the International Conference on Computational Science and Computational Intelligence (CSCI)*, 15-17 March 2017, Las Vegas, NV, USA, Online ISBN: 978-1-5090-5511-1, E-ISBN: 978-1-5090-5510-4, DOI: 10.1109/CSCI.2016.0217, pp. 62-73, Published by IEEE, Available: https://ieeexplore.ieee.org/document/7881510.

[9] Marsida Toska, Joakim Nivre and Daniel Zeman, "Universal Dependencies for Albanian", in *Proceedings of the Fourth Workshop on Universal Dependencies,* 13 December 2020, Bracelona, Spain, Online ISBN 978-1-952148-48-4, pp. 178-188, Published by Association for Computational Linguistics, Available: https://www.aclweb.org/anthology/2020.udw-1.20 .

[10] Nelda Kote, Marenglen Biba and Evis Trandafili, "A Thorough Experimental Evaluation of Algorithms for Opinion Mining in Albanian", in *Proceedings of the International Conference on Emerging Internetworking, Data & Web Technologies,* Lecture Notes on Data Engineering and Communications Technologies, February 2018, Thailand, Online ISBN: 978-3-319-75927-2, E-ISBN: 978-3-319-75928-9, DOI: 10.1007/978-3-319-75928-9_47, pp. 525-536, Published by Springer, Cham, Available: https://link.springer.com/chapter/10.1007/978-3-319-75928-9_47.

[11] Daniel Vasić and Emil Brajković, "Development and Evaluation of Word Embeddings for Morphologically Rich Languages", in *Proceedings of the 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 13-15 September 2018, Split, Croatia, Online ISBN: 978-9-5329-0087-3, E-ISBN:978-1-5386-6770-5, DOI: 10.23919/SOFTCOM.2018.8555822, pp. 1-5, Published by IEEE, Available: https://ieeexplore.ieee.org/document/8555822.

[12] Nelda Kote, Marenglen Biba and Evis Trandafili, "An Experimental Evaluation of Algorithms for Opinion Mining in Multi-domain Corpus in Albanian", in *Conference Proceedings of the International Symposium on Methodologies for Intelligent Systems – Foundations of Intelligent Systems*, 29-31 October 2018, Cyprus, Limassol, Cyprus, Online ISBN: 978-3-030-01851-1, DOI: 10.1007/978-3-030-01851-1_42, pp. 439-447, Published by Springer, Available: https://link.springer.com/chapter/10.1007/978-3-030-01851-1_42.

[13] Evis Trandafili, Nelda Kote and Marenglen Biba, "Performance Evaluation of Text Categorization Algorithms Using an Albanian Corpus", in *Proceedings of the International Conference on Emerging Internetworking, Data & Web Technologies – Advances in Internet, Data & Web Technologies,* Lecture Notes on Data Engineering and Communications Technologies, February 2018, Thailand, Online ISBN: 978-3-319-75927-2, E-ISBN: 978-3-319-75928-9, DOI:10.1007/978-3-319-75928-9_48, pp. 537-547, Published by: Springer Cham, Available: https://link.springer.com/chapter/10.1007/978-3-319-75928-9_48.

[14] Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Geraldine Walther, Ekaterina Vylomova *et al.*, "UniMorph 2.0: Universal Morphology", in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* May 2018, Miyazaki, Japan, Online ISBN 979-10-95546-00-9, pp. 55-61, 7-12 Published by European Language Resources Association (ELRA), Available: https://www.aclweb.org/anthology/L18-1293.

[15] Besim Kabashi and Thomas Proisl, "A proposal for a part-of-speech tagset for the Albanian language", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* 23-28 May 2018, Portoroz, Slovenia, Online ISBN: 978-2-9517408-9-1, pp. 4305-4310, Published by European Language Resources Association (ELRA), Available: http://www.lrec-conf.org/proceedings/lrec2016/summaries/1066.html.

[16] Indrashis Das, Bharat Sharma, Siddharth S. Rautaray and Manjusha Pandey, "An Examination System Automation Using Natural Language Processing", in *Proceedings of the 2019 International Conference on Communication and Electronics Systems (ICCES)*, 17-19 July 2019, Coimbatore, India, Online ISBN:978-1-7281-1261-9, E-ISBN: 978-1-7281-1262-6, DOI: 10.1109/ICCES45898.2019.9002048, pp. 1064-1069, Publishted by IEEE, Available: https://ieeexplore.ieee.org/document/9002048.

[17] Aditya Jain, Gandhar Kulkarni and Vraj Shah, "Natural Language Processing" in *International Journal of Computer Sciences and Engineering*, Print ISSN: 2347-2693, pp. 161-167, vol. 6, Issue 1, January 2018, DOI: 10.26438/ijcse/v6i1.161167, Available: https://www.ijcseonline.org/full_paper_view.php?paper_id=1652.

[18] Dongyang Wang, Junli Su and Hongbin Yu, "Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language", in *IEEE Access*, 14 February 2020, Online ISBN: 2169-3536, DOI: 10.1109/ACCESS.2020.2974101, pp. 46335–46345, Published by IEEE, Available: https://ieeexplore.ieee.org/document/8999624.

[19] Ritu Banga and Pulkit Mehndiratta, "Tagging Efficiency Analysis on Part of Speech Taggers", in *Proceedings of the 2017 International Conference on Information Technology (ICIT),* 23 December 2018, Bhubaneswar, India, Online ISBN 978-1-5386-2924-6, E-ISBN: 978-1-5386-2925-3, DOI: 10.1109/ICIT.2017.57, pp 56-62, Publisheed by IEEE, Available: https://ieeexplore.ieee.org/abstract/document/8423919.

[20] Rushali Dhumal Deshmukh and Arvind Kiwelekar, "Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing", in *Proceedings of the 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA),* 5-7 March 2020, Bangalore, India, Online ISBN 978-1-7281-4167-1, E-ISBN: 978-1-7281-4168-8, DOI: 10.1109/ICIMIA48430.2020.9074941, pp. 50-62, Published by IEEE, Available: https://ieeexplore.ieee.org/document/9074941.