IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 21, 2019, accepted December 24, 2019, date of publication January 1, 2020, date of current version January 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963674

ADD: Academic Disciplines Detector Based on Wikipedia

ANA GJORGJEVIKJ, KOSTADIN MISHEV, AND DIMITAR TRAJANOV
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, 1000 Skopje, Macedonia

Corresponding author: Ana Gjorgjevikj (gjorgjevikj.ana@students.finki.ukim.mk)

This work was supported in part by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

ABSTRACT The academic disciplines and their interrelationships represent a backbone that organizes the enormous amount of documented human knowledge available today. Having an up-to-date overview of the established disciplines, the emerging ones, and their mutual interactions is essential to the academic institutions, publishers, and many other actors involved in today's knowledge-based society, even in a situation of nonexistence of a precise definition of the term ''academic discipline'' itself. The discipline classification schemes represent crucial resources for the purpose, and in circumstances where the knowledge production rate demands discovering changes in their structure very frequently, the data-driven methodologies which facilitate their revision processes become essential. Analyzing the world-wide community's opinion on what represents a discipline, available through Wikipedia, can be very informative for the purpose, considering Wikipedia's comprehensiveness, continuous updates, and historical exports availability. This paper proposes a data-driven methodology for identification of the concepts which the world-wide community defines as disciplines at a particular moment by analyzing the information available in Wikipedia at that same moment. At the same time, it discusses Wikipedia's strengths and challenges on the task while also comparing a variety of Machine Learning and Natural Language Processing methodologies. High accuracy of the trained models is achieved on datasets created for this task specifically, and low changes in the model accuracy are observed on four Wikipedia exports from 2015 to 2018.

INDEX TERMS Machine learning algorithms, natural language processing, academic discipline, text analysis, Wikipedia.

I. INTRODUCTION

The philosophical debate on the nature of the academic disciplines, their status in modern society, and their future development is long-lasting and comprehensive. The perception of the concept of ''discipline'' and the classical disciplinary organization of human knowledge has been changing through time as a natural consequence of the scientific and technological advancements, as well as changes in the context in which new knowledge is being produced, shifting from disciplinary towards broader, application-oriented context [1]. On the contrary to the initial, mainly archival, function of the disciplines, today they have complex and at the same time essential roles in the modern society, like being primary units of structure and differentiation in the modern system of science, structures vital to tracking the scientific development, subjects taught in schools, means for designation of professional occupations, active and mutually interacting knowledge-producing systems [2].

The organization of the disciplines into classification systems, particularly in hierarchical ones where one discipline is defined as a parent to others, is another complex task, even though the modern classification systems are more dynamic and open to the vague and continually changing boundaries among the disciplines, compared to the closed and finite systems from the past [3]. Besides the unquestionable difficulty of updating the existing classification systems in conditions of unclear rules for disciplines differentiation and emerging disciplines detection, the necessity to accurately capture the current state of human knowledge and research activity, as well as to fit the emerging and rapidly changing scientific concepts/theories, create constant interest in automating the detection of new/obsolete disciplines and changes in their interrelations.

Contrary to the conceptions of an intellectual inferiority of the crowds compared to single individuals' intelligence,

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

the theories supporting the "wisdom of the crowd" emphasize the idea that under the right circumstances, groups can be remarkably intelligent and sometimes smarter than the smartest individuals in those groups. This phenomenon indicates that a large group of people can be smarter than an elite few when there is diversity, independence in the opinions and decentralization, so opinion aggregation results in judgments closer to the truth than those of a single or few individuals [4]. Most of these theories, however, do not deny the importance of the individual's expertise, indicating that the presence of knowledgeable and well-informed individuals can improve the group's collective judgments. The wisdom of the crowd and the idea of dispersed knowledge between a large number of individuals in the society are among the essential concepts behind Wikipedia's distributed and unrestricted editing policy. Wikipedia[1] is an example of successful collaboration between a large number of editors, with an observed shift of the edits distribution from an "elite" group of users (in its beginnings) towards the "common" users [5], resulting in a resource with accuracy comparable to traditional encyclopedias and professional textbooks in certain domains [6]–[8], resilient to malicious editing through distributed monitoring [9], [10]. It is evident that Wikipedia's articles vary in quality, but in some cases, it is only a misperception that a particular Wikipedia article is less credible than an article on the same topic in a traditional encyclopedia [11]. The precisely defined article life cycle[2] further helps in distinguishing articles credibility. A question for debate, however, is whether the value of Wikipedia should be measured by comparing it to traditional encyclopedias, considering its wide number of unique characteristics, like being a source of information not covered by traditional encyclopedias (e.g., entertainment), as well as its scientific actuality [12].

Motivated by the challenges related to tracking of the disciplines evolution through time and timely detection of emerging disciplines, this work studies Wikipedia's potential in addressing these challenges and complementing the current state-of-the-art methodologies based on scientific publications analyses. Besides analyzing how Wikipedia's continuous evolution, comprehensiveness, and historical exports availability can contribute to solving the challenges, the second contribution is comparing and evaluating the novel methodologies from the domains of Natural Language Processing (NLP), Machine Learning and its sub-fields Deep Learning and Transfer Learning on the large amount of data available in Wikipedia, underutilized in the domain of interest, and discussing the challenges of the process. This work proposes a methodology for detection of the concepts which Wikipedia's community has defined as disciplines at certain point of time and separation of those that exhibit similar features to the disciplines already part of classification schemes, focusing on "academic discipline" as thematic structure of interest (interchangeably referred to as "discipline" through the paper). The implemented methodology, i.e., Academic Disciplines Detector (ADD), is evaluated on four Wikipedia exports between 2015 and 2018 and available on GitHub[3].

II. RELATED WORK

This section starts with a comparison of several widely used classification systems related to academic disciplines and programs. While some primarily serve for classification of disciplines, academic programs (e.g., Classification of Instructional Programs (CIP)) or fields of science (e.g., Field of Science and Technology (FOS) Classification), some incorporate the disciplines into their structure, but primarily serve for library resources organization and retrieval (e.g., Dewey Decimal Classification (DDC), Universal Decimal Classification (UDC), and Library of Congress Classification (LCC)). CIP is a taxonomy of instructional programs developed to facilitate the process of organization, collection, and reporting of fields of study or program completion. Initially created in 1980 and subjected to revisions in 1985, 1990, 2000, 2010, with the following one in 2020, its hierarchy consists of three levels, the top one representing a general grouping of related programs, the intermediate one a grouping of programs with comparable content and objectives, and the lowest one representing specific instructional programs [13]. FOS is a science classification system intended for R&D expenditure of the government, higher education, private non-profit, and business enterprise sector. It is a three-level schema with six major fields at the broadest level [14]. DDC is a continuously revised classification system, providing a general organization of knowledge. While it is primarily used as a library classification system, it can also support web resources organization and retrieval. At the broadest level, it is divided into ten main classes, further divided into ten divisions and ten sections [15]. UDC has been created as a detailed, flexible indexing language for information retrieval [16]. Its analytico-synthetic nature allows the expression of an unlimited combination of subject attributes and relationships. It has a discipline-based organization of knowledge, i.e., concepts are placed under the field which studies them. The hierarchic structure has ten classes, subdivided into their logical parts, where each subdivision is further subdivided as needed [17]. LCC is one of the world's most widely adopted library classification systems, initially developed to organize the book collections of the Library of Congress. The top-level classes represent the main fields of human knowledge (academic disciplines), divided into sub-classes that represent their brunches. The further subdivisions correspond to form, place, time, or topical aspects [18].

A large number of research papers propose automated methodologies for identification of thematic structures in science. In a brief overview of part of them, we specifically focus on the definition of the thematic structures they identify and their methodology. Waltman and Van Eck [19] propose

[1] https://www.wikipedia.org/

[2] https://en.wikipedia.org/wiki/Wikipedia:Article_development

[3] https://github.com/f-data/ADD

a three-step methodology for constructing a three-level hierarchical classification system of science, which starts with determination of the scientific publications relatedness based on their direct citations, then clustering and cluster labeling. From publications in Web of Science (WoS) for the period of 2001-2010, classification with 20 research areas at the top level, 672 in the middle, and 22,412 at the lowest level is extracted. The same methodology is used in [20] and [21], in the later combined with direct citation-based methodology [22] to detect emerging topics in science and technology from citation databases. Salatino *et al.* [23] present the Computer Science Ontology, an automatically generated ontology of research areas with approximately 26K topics and 226K semantic relationships. The algorithm, Klink-2 [24], takes as input a set of scholarly keywords and their relationships with other entities (research papers, venues, authors, organizations) to output an ontology with semantic relationships between the research topics identified in the provided keywords and data. Using the Klink-2 algorithm to detect research areas, Salatino *et al.* [25] show that the emergence of a scientific topic can be detected in an ''embryonic'' phase when weakly connected research areas start to interact. Suominen and Toivanen [26] analyze scientific publications with topic modeling. Using publications from authors with Finnish affiliation, 60 topics clustered in 5 communities are identified and interpreted with regard to two expert-created classifications. Leydesdorff and Rafols [27] and Leydesdorff *et al.* [28] identify the disciplinary structure of science on a macro level based on factor analyses of the journal citation matrices and the WoS subject categories. The factors are interpreted in terms of disciplines. Meng *et al.* [29] detect scientific disciplines through affinity propagation clustering of the WoS indexed journals, represented through TF-IDF vectors based on their textual descriptions and vectors based on journals cross-citations, whose similarity is determined as vector angle cosine. Instead of citations, Bollen *et al.* [30] use clickstream data from scholarly web portals. The user interactions, aggregated at journal level to calculate a journal transition probability, are used to generate science maps. Chavalarias and Cointet [31] represent scientific fields through key-phrases in publications, connected in a co-word network, and clustered using clique percolation method. They analyze the changes in the network with a goal to reconstruct the cognitive evolution of science. Herrera *et al.* [32] represent scientific fields by use of a community finding algorithm in a scientific concepts network, where two concepts are linked if they appear in the same publication. They study the evolution of scientific fields through network-based analysis.

A smaller number of papers study Wikipedia's potential in identification of thematic structures in science. Salah *et al.* [33] compare the differences between the UDC and Wikipedia category structure under the *Arts* category. Wikipedia category membership data is used to assign a level to each category under the *Main topic classification* category. Then the *Arts* category and its related categories are mapped to UDC and compared for their structure. Minguillón *et al.* [34] present a semi-automatic method based on random walks to determine a subset of Wikipedia articles containing scientific and technological content. 60,108 Spanish Wikipedia pages in 340 communities were identified as containing scientific and technological content, reachable from 974 six-digit categories from the UNESCO nomenclature for fields of science and technology. Joorabchi and Mahdi [35] present a methodology for indexing library metadata records with Wikipedia concepts, where the Wikipedia concepts appearing in the library metadata are identified and classified as ''key'' or ''non-key'' based on 15 statistical, positional, and semantic features. The training and test datasets consist of a manually indexed subset of the WorldCat-Million dataset (DDC class 006.3, Artificial Intelligence) with Wikipedia concepts, where 469 Wikipedia concepts are labeled as ''key'' and 1,293 as ''non-key'' concepts. In a more recent paper, Joorabchi and Mahdi [36] present methodology and software system for automatic mapping of Faceted Application of Subject Terminology (FAST) subject headings, a controlled vocabulary based on the Library of Congress Subject Headings (LCSH), to their corresponding Wikipedia articles. A binary classifier is trained to classify the candidate Wikipedia articles, represented by a set of 14 positional, statistical, and semantic features, into a ''corresponding'' or ''non-corresponding'' class. The dataset used by the algorithm contains 170 FAST subject headings manually mapped to their corresponding Wikipedia articles. Yoon *et al.* [37] construct a classification scheme of science and technology by extracting its backbone from Wikipedia, using the nodes reachable from the *Scientific disciplines* category. To extract the backbone of the network, pruning of insignificant links using the shortest path information and reduction using local structural information is done. Korean Wikipedia dump from 2017-09-09 and All Science Journal Classification (for validation) are used.

While most of the related work in the field is based on lexical, citation-based, or hybrid analysis of scientific publications in electronic publication databases, this work attempts to solve the discipline detection task by exploiting the knowledge available in a community created encyclopedia. The thematic structure this work detects is ''academic discipline,'' while most of the related work detects lower-level thematic structures such as scientific/research ''topics''/''terms'', resulting in the detection of a larger number of ''units'' compared to our method. On the contrary to the common use of unsupervised machine learning methods, this work is based on supervised methods, incorporating the ''ground truth'' knowledge from an expert classification scheme into the training/test data. Most of the related work based on Wikipedia utilizes the article interlinks or the category graph in conjunction with network analyses to identify articles/categories referring to disciplines or scientific concepts [33], [34], [37]. Those that use machine learning algorithms to classify Wikipedia articles as ''appropriate'' or not in a specific context, train their models on a smaller number of manually engineered features and smaller datasets compared to the method presented in this work, which in its core module uses automatically extracted features of larger dimension and larger training/test datasets. This work further evaluates the applicability of the novelties from the domain of Natural Language Processing, as well as from Machine Learning sub-fields Deep Learning and Transfer Learning in solving the task, which are not used in the referenced related work. Furthermore, it compares a large number of methods/configurations and studies the final methodology accuracy change over time on several Wikipedia exports.

III. DATA

The data used in this research comes from two primary data sources, Wikipedia and the CIP classification schema.

The data available through the public Wikipedia XML exports is utilized throughout this work, or, more precisely, the XML files containing the current revisions of all English Wikipedia articles, titled as *enwiki-YYYYMMDD-pages-articles.xml*. The latest files may be obtained from the Wikimedia dump directory,[4] while the older ones from other available sources.[5] Throughout the methodology development and evaluation phase, the export from 2017-06-01 was used. For evaluation of the changes in the results accuracy through time, three additional exports from 2015-06-02, 2016-06-01 and 2018-11-20 were used.

During the development phase, data (title and description) from an expert-created classification schema, CIP version from 2010, was utilized. CIP offers a short description of the objective and instructional content of the academic programs [13], making it very convenient for use in this work, first, for defining the ''ground truth'' for evaluation, second for understanding the common linguistic patterns which the academic disciplines/programs titles and descriptions follow and which influenced individual decisions in the methodology development phase.

For representation of the textual data into fixed-length vector form using pre-trained text encoding models, in addition to the model files, word vectors trained with GloVe [38] on Common Crawl (840B tokens)[6] and with FastText [39], again trained on Common Crawl (2 million word vectors)[7] were utilized as they were part of the requirements of some of the text encoding models. The details are available in Section IV-F.2.

IV. METHODS

A. SYSTEM ARCHITECTURE

This work attempts to solve the problem of detecting the disciplines which are current at a particular moment through a multistage processing pipeline working over the English Wikipedia content from that same moment. The availability of historical Wikipedia exports gives an opportunity for evaluation of the proposed methodology on the most current, as well as older data, and analyzing the changes in the results through time. Fig. 1 illustrates the architecture of the implemented system and the data flow between its main modules. The system accepts XML export file(s) with the current revisions of all Wikipedia articles as input and starts with the extraction of articles metadata, text, and list of all articles linked in the analyzed article's content. The article metadata and text are passed to the Basic Filter, which implements filtering heuristics to quickly determine if an article potentially refers to a discipline or not, preventing unnecessary processing of huge data volumes by the subsequent modules. The Lead Section Excerpts Extractor then extracts short, representative excerpts from the retained articles lead section, as further described in Section IV-E. Wikipedia articles lead section appears before the table of contents and articles first heading[8] to provide a concise overview of the article's topic, i.e., a definition/identification of the topic and a summarization of the important points, with the first several sentences describing the article's topic notability. Depending on the article length, the lead section varies between one and four paragraphs. All lead section excerpts of potential disciplines are passed to the Text Classifier, which implements text-based classification logic to detect all articles which contain a definition of a discipline in their lead section. Nevertheless, not all Wikipedia articles which resemble a discipline or contain a discipline definition in their lead section first sentences refer to an actual discipline (discussed into more details in Section IV-G) or are mature enough to enter classification schemes. Therefore a final module is added to make a distinction between them. The Node Classifier is trained to recognize the discipline candidates that have similar features to those already present in classification schemes by combining the results from the previous module with candidates' graph-based features. The application modules are implemented in Python, using NLTK [40] and Gensim [41] for text processing tasks, Scikit-learn [42] for machine learning tasks and NetworkX [43] for graph analyses. For utilization of pretrained text encoding models, TensorFlow,[9] TensorFlow Hub[10] and PyTorch[11] were used.

The non-existence of clear ''ground truth'' on what represents a discipline and what does not, due to the difficulty of defining objective and precise boundaries between the disciplines, makes the discipline detection methodologies hard to be entirely objectively evaluated. To come to a method that is valid and objective to the extent to which it can be evaluated as such, as well as to avoid bias towards any possible ''incorrect'' stance encoded in Wikipedia, the viewpoints of the subject experts encoded in an existing classification scheme of disciplines, CIP, were taken as ''ground truth'' in creation of the training and evaluation datasets used by the

[4] https://dumps.wikimedia.org/

[5] https://meta.wikimedia.org/wiki/Data_dumps#Download

[6] https://nlp.stanford.edu/projects/glove/

[7] https://fasttext.cc/docs/en/english-vectors.html

[8] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

[9] https://www.tensorflow.org/

[10] https://www.tensorflow.org/hub

[11] https://pytorch.org

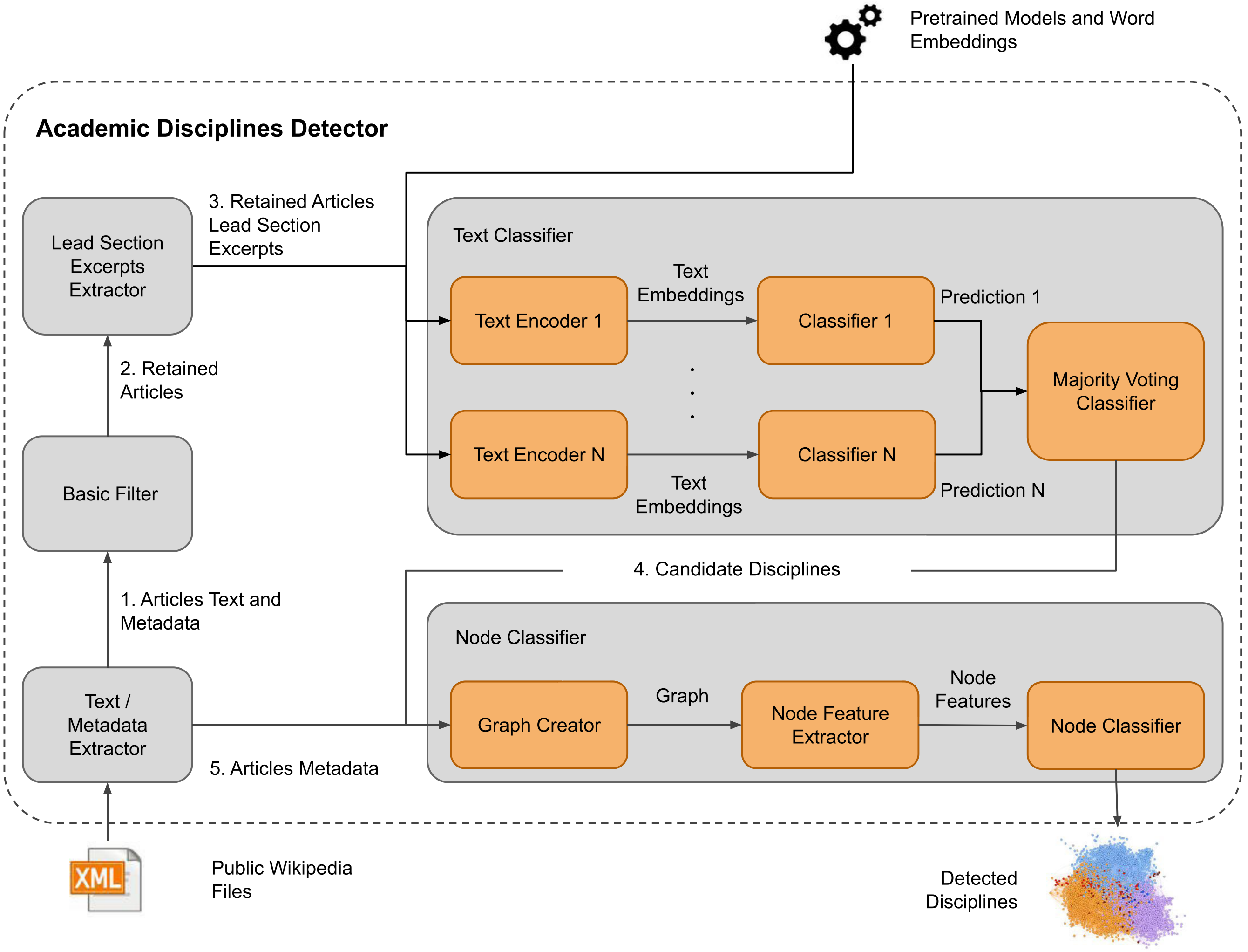

FIGURE 1. **High-level system architecture.**

application modules. Considering the complex philosophical debates on the subordination of the disciplines with regard to one another, this work does not attempt to infer or suggest the hierarchical ordering of the detected disciplines.

B. COMPARISON OF WIKIPEDIA AND CIP

This section examines the alignment of the lead section excerpts of articles that refer to disciplines in Wikipedia, extracted as described in Section IV-E, and the short description of the disciplines offered in CIP. CIP is also convenient for understanding the common linguistic patterns in the discipline titles and descriptions. Studying the similarity/difference between the two datasets allows their proper combination to maximize the accuracy on the task.

To get an understanding of the most common, less common, or non-appearing word category types in the discipline titles, important for the development of a basic filtering heuristics, all titles of the CIP classes were subjected to part-of-speech (POS) tagging and frequency distribution analysis. It allows performing an initial filtering of the Wikipedia articles that almost certainly do not refer to a discipline only by analyzing their title. Since all words in the CIP titles are capitalized, and incorrect capitalization negatively influences the POS tagging accuracy, true case identification and manual correction of the capitalization was done before the POS tagging. The frequency of each Penn Treebank POS tag is given in Fig. 2 (a), whereas the most common POS tag n-grams are given in Fig. 2 (b). Fig. 2 (a) indicates that the most common POS tags are the singular/mass nouns (NN), adjectives (JJ), plural nouns (NNS) and conjunctions (CC). The proper nouns (NNP and NNPS) and the other word forms appear rarely. Fig. 2 (b) shows that the most common n-grams again consist of a combination of general nouns, adjectives, and conjunctions. Patterns consisting of proper nouns only, where all title words would be capitalized, are not among the frequent ones. These findings are in line with our expectations that the name of a discipline normally consists of general nouns, adjectives, conjunctions, and less frequently of verbs, proper nouns, cardinal numbers, further elaborated in Section IV-D.

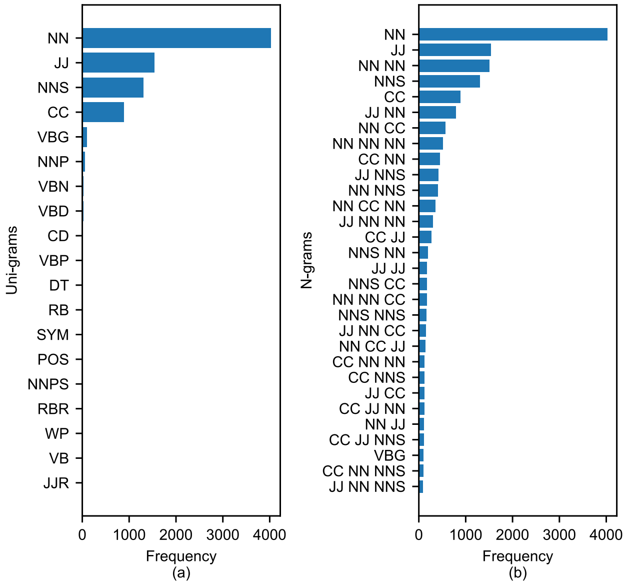

FIGURE 2. Frequency distribution of Penn Treebank POS tagset in CIP titles. (a) Uni-grams. (b) Uni-grams, bi-grams and three-grams.

To get an understanding of the lead section in Wikipedia articles referring to disciplines, a comparison with the descriptions of the disciplines in CIP was made. The main concern was the possible writing style difference between the formal schemes and Wikipedia, in which case, using the descriptions only from the formal classification scheme as positive samples in the training and test dataset will result in an unrepresentative dataset and a model that might perform poorly on the "real" samples. The second concern was the selection of an appropriate lead section excerpts length, considering the differences in the lead section length between different Wikipedia articles. To create a set of disciplines whose descriptions will be compared, a simple matching between the CIP titles and the Wikipedia article titles was done, preceded by a process of lemmatization. More complex string matching techniques were out of the scope of this task since its purpose was not to find all mappings, but to evaluate the description similarity of the classes/articles that almost certainly refer to the same discipline. For the representation of the class descriptions and articles lead section excerpts in a vector space, the Bag of Words (BoW) model [44] with Term Frequency - Inverse Document Frequency (TF-IDF) weighting was used, followed by a cosine similarity calculation. The created dataset consisted of around 150 mappings and Fig. 3 (a) shows that the distribution of the compared descriptions/excerpts lengths is similar. The texts were subjected to preprocessing consisting of tokenization, stop-word removal, and lemmatization, before being represented as vectors in a common n-dimensional vector space $\mathbb{R}^n$ where each dimension is associated with one vocabulary (V) term $t_i \in V$, $i \in [1, n]$. In each dimension i, the vector component resembled the TF-IDF weight of the appropriate term t_i, and the text being represented as a vector. Each matched discipline was assigned two vectors, one for its CIP description and one for its Wikipedia lead section excerpt, the similarity of which was calculated as a cosine of their angle. The distribution of the cosine similarities between the CIP description and the Wikipedia lead section excerpt for each matched discipline is given in Fig. 3 (b), with mean cosine similarity equal to 0.27. Additional analyses of CIP and Wikipedia in the context of educational content classification are available in [45].

C. TEXT/METADATA EXTRACTOR

The Text and Metadata Extractor module processes raw XML files containing the current revisions of the Wikipedia articles to extract each article's metadata, text and list of articles it references in the content. To a large extent, the module relies on the parsers available in Gensim to extract the metadata and articles text cleaned from markup, to recognize and filter the articles not belonging to the main namespace (namespace containing encyclopedia articles, lists, disambiguation pages and redirects[12]) and stub articles with less than 200 characters (the default suggested number of characters in the parser). To improve the quality of the extracted data for the task of interest, prior to the extraction using Gensim, an optional, custom markup cleaning step was introduced, dependent on the particular Wikipedia XML export files.

D. BASIC FILTER

Wikipedia's community has created a broad set of guidelines and policies that help editors identify the best practices for the type of article they are creating or editing. The one important for this paper is the way article titles are constructed[13]. The titles of the articles are written in sentence case, with the initial letter capitalized by default. Words in the title are not capitalized unless they are capitalized in a normal text. The nouns in the titles are in a singular form, except for the nouns that only have a plural form, no indefinite articles are placed at the beginning of the titles, except when they are part of proper names. Nouns and noun phrases are preferred in titles over other part-of-speech forms.

Additionally, the presented analyses of the word types in CIP titles showed that they mostly consist of general nouns, almost never of proper nouns only, and rarely contain numbers (Fig. 2). Based on the findings above, several general rules on the titles of Wikipedia articles that refer to disciplines can be constructed, under the assumption that these articles can be considered policy compliant. These rules, outlined below, on their own, are insufficient to detect all the articles referring to disciplines.

One-word discipline titles are expected to be general nouns, with the first letter capitalized solely. Therefore, the article titles that consist of one word only and are written in all uppercase or contain digits are filtered. Multi-word discipline titles are again expected to contain at least one general noun and no digits, so accordingly, multi-word article

[12] https://en.wikipedia.org/wiki/Wikipedia:Namespace

[13] https://en.wikipedia.org/wiki/Wikipedia:Article_titles

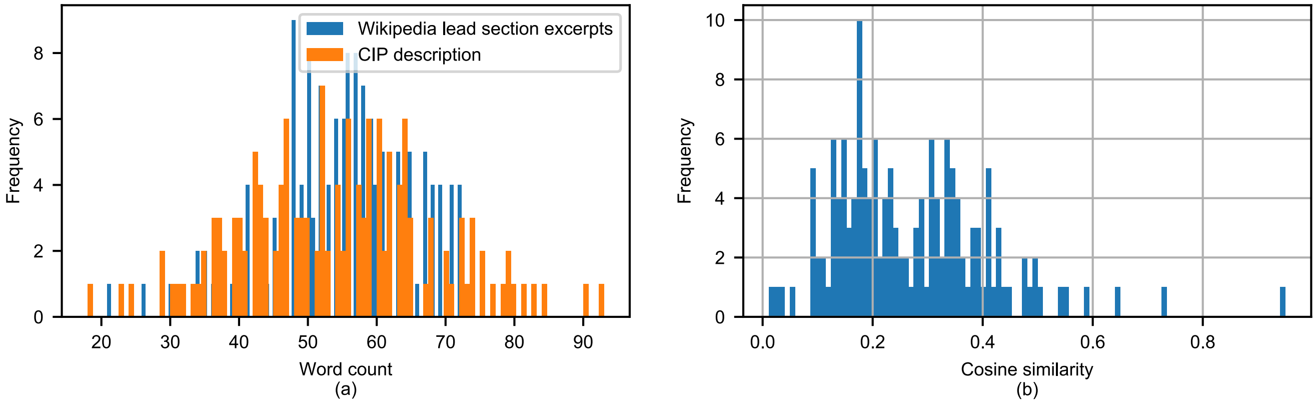

FIGURE 3. Comparison of the matched disciplines lead section excerpts and CIP descriptions. (a) Word count distribution. (b) Cosine similarity distribution.

titles with all words capitalized, containing digit, referring to an admin or disambiguation page are filtered. The admin pages are identified based on a set of particular words and phrases that are common for certain types of admin pages (lists, time-lines, indexes, glossaries, outlines, and other). The disambiguation pages are identified based on the common word they contain in the title, as well as by common phrases appearing in the article text.[14] When applying the filtering to all available English articles, less than 1/3 remain for further analyses, and the statistics by processed Wikipedia export is presented in Section V. Only this subset of articles is taken as relevant for the task and utilized when generating training and test datasets in the subsequent modules.

E. LEAD SECTION EXCERPTS EXTRACTOR

The extracted lead section of each Wikipedia article retained by the Basic Filter is subjected to further cleaning and processing to extract a smaller excerpt that ensures a word count distribution of the excerpts similar to the one of the CIP descriptions, as well as a concise and sufficiently distinctive definition of the concept/entity the article refers to. The excerpt extracted by this module consists of the first sentences from the article lead section, appended until an approximate mean word count of 50(±25) is reached, as applicable. The article title is then appended as a first sentence of the lead section excerpt.

F. TEXT CLASSIFIER

Considering Wikipedia's comprehensiveness and less strict revision process compared to the expert-created classification schemes, we assume that a slightly larger number of disciplines may be defined by the Wikipedia's community as such than present in expert-created classification schemes at specific moment. Nevertheless, considering the number of Wikipedia articles retained by the Basic Filter, even in a less restrictive setting, we do not expect that more than 1% of these Wikipedia articles refer to disciplines. Since the number of articles that refer to disciplines is substantially smaller than the number of those that do not, the data which the Text Classifier deals with is highly imbalanced. At this stage, the problem of detecting articles defined as disciplines by the Wikipedia community is formalized as binary classification of an imbalanced dataset, composed of lead section excerpts of articles referring to disciplines (positive samples) and excerpts from articles dealing with any other topic, including science (negative samples). The subsequent sections provide details on the training/test dataset creation, the representation of the textual excerpts as fixed-length vectors, and the classification process.

1) TRAINING/TEST DATASET CREATION

The supervised learning algorithms require the existence of labeled training and test datasets, and in line with the problem formulation, the datasets have to be composed of short textual paragraphs labeled as either positive or negative samples. The "real world" data which the trained classifier will be working with is coming from Wikipedia and the differences between the disciplines descriptions in CIP and Wikipedia, presented in Section IV-B, indicated usage of different wording in both, leading to a conclusion that basing the positive samples on CIP descriptions solely would not create a representative dataset of the real Wikipedia data. To still utilize CIP in enlargement of the number of positive samples, two training datasets were created based on CIP and Wikipedia using several different selection criteria described below.

a: POSITIVE SAMPLES SELECTION

The candidates for positive samples were selected from a set of Wikipedia articles mapped to disciplines present in CIP and from several Wikipedia articles explicitly listing disciplines. At this stage, the mapping between Wikipedia and CIP was done by an approximate matching between the Wikipedia article titles and the CIP titles, followed by a manual validation. The second source of positive samples was

[14] https://en.wikipedia.org/wiki/Wikipedia:Disambiguation

the article *Outline of academic disciplines*[15] which provides an overview and a hierarchical arrangement of academic disciplines linked to their Wikipedia articles. Extraction of the linked articles and manual revision was done due to the presence of links to articles which, in our opinion, do not explicitly define an academic discipline (e.g., *Piano, Black hole*). The category *Academic disciplines*,[16] which groups articles and other categories related to academic disciplines, was used as a third source of samples. The articles that belong to this category or to one of its sub-categories up to four levels down in the hierarchy were extracted as potential positive samples. Manual revision of the extracted samples was done again, labeling each sample as either positive or negative, again due to the potential presence of incorrectly categorized articles. Finally, additional selection heuristic was applied to all Wikipedia articles by using common terms that appear in disciplines names (e.g., *study, discipline*), or names of established disciplines (e.g., *Engineering, Science, History, Philosophy*) to extract additional candidates, again subjected to a manual revision. The available CIP classes were subjected to filtering of the deleted/moved ones and used as positive samples.

b: NEGATIVE SAMPLES SELECTION

The candidates for negative samples were selected through several methods. The first one relied on Wikipedia categories which group articles that refer to named entities, i.e., people, locations, organizations, products. It traversed the hierarchy of sub-categories and articles under categories such as *Companies by industry, Film actors by nationality, Singers by nationality, Sport teams by sport, Rivers by country, Video games by genre* and many other, up to six levels of children. The second source of negative samples were articles that describe scientific terminology and which are very closely related to the academic disciplines, but not describing an actual one. Those types of samples are considered important since the articles very closely related to science most probably use similar wording in their text to the actual discipline articles, and it would be crucial for the classifier to learn to distinguish them. To extract terminology related to science, approximately thirty glossary articles under the category *Glossaries of science*[17] were processed to find the actual articles linked from the glossary pages. The third source of candidates was the hierarchy of categories like *Scientists by century, Philosophers by field* and similar, grouping articles for people closely related to science, again probably sharing much of the wording with the disciplines articles. To come to a sufficiently large number of negative samples for our dataset, the rest of the negative samples were selected randomly from the whole filtered Wikipedia dataset. The candidate negative samples were validated not to contain terminology common for academic disciplines in their titles and lead section excerpts. A validation heuristic was inevitable due to the training dataset size, i.e., approximately 83,000 samples, making the manual validation impossible.

Two training datasets were generated, one consisting of Wikipedia samples only and one as a combination of samples from both Wikipedia and CIP, where CIP provides positive samples only. The test dataset was created by randomly selecting 5,000 Wikipedia articles and manually labeling them as positive or negative. The ratio of the positive and negative samples in the test set is 1% positive and 99% negative, making it a representative sample of the Wikipedia article distribution hypothesized at the beginning of this section.

2) TEXT REPRESENTATION

One of the prerequisites for using machine learning for text classification is the representation of the variable-length texts into a machine-understandable format, namely fixed-length feature vectors, and the representation quality has as large influence on the classification algorithm performance as the classifier selection. The most widely used and still influential methods for text representation are the ones based on bag of words (BoW) or bag of n-grams (BoN), where, as the name suggests, the order of the words/n-grams in the text is not taken into account when deriving the vector representation. The resulting vectors are usually high-dimensional and sparse because each feature (dimension) corresponds to one word (or n-gram) in the vocabulary, created from the whole text corpus. The vector dimensionality and sparsity, the ignoring of the word order, and semantics are commonly pointed out as major shortcomings of the BoW method, but its simplicity for understanding, its efficiency, and acceptable accuracy on many tasks are still unarguable. Strategies for addressing the shortcomings and improving the performance of the basic BoW model are available, like considering n-grams instead of single words to take into account the word order in short contexts, preprocessing with stop words removal, stemming or lemmatization methods, feature selection (retaining the features most correlated with the target variable or features with highest variance) or projection to low dimensional space to reduce the dimensionality. To capture information of words/n-grams co-occurrence in the text, weighting methods such as TF-IDF and its variations can be used. With the success of the deep neural network architectures in many areas, especially in learning general, dense word representations (embeddings), the task of learning general and dense vector representations of sentences, paragraphs, and documents using neural networks gained increased popularity. The purpose of learning general embeddings of text is the transferability, namely, reusing the pre-trained embeddings or models on some general dataset in many other domain-specific tasks that might lack sufficient training data. The most simple method for generating text vectors is averaging previously learned word embeddings on large, general-domain text collections (used as they are or fine-tuned). Some of the more advanced methods for learning general text representations are

[15] https://en.wikipedia.org/wiki/Outline_of_academic_disciplines
[16] https://en.wikipedia.org/wiki/Category:Academic_disciplines
[17] https://en.wikipedia.org/wiki/Category:Glossaries_of_science

TABLE 1. Dimensions of the embeddings produced by the used text encoders. The BoN based encoder, fitted on the training datasets, produces embeddings with different dimensions for the two datasets, depending on the vocabulary size and if preprocessing is used. The BoN based encoder variations retaining top *k* features are not displayed in the table.

	BoN (All features, P-value<0.1)		Universal Sentence Encoder	InferSent	
Dataset	Preprocessing - Yes	Preprocssing - No	Transformer	GloVe Word Embed.	FastText Word Embed.
Wikipedia	3,395	3,481	512	4,096	
Wikipedia & CIP	5,439	5,500	512	4,096	

Paragraph Vectors [46], Skip-Thought [47], FastSent [48], InferSent [49], Universal Sentence Encoder [50] and others.

Selecting a method for representing text as a fixed-length vector is a non-trivial task, and no recommended best practice exists. The plentitude of available methods leads to the conclusion that the best one for domain-specific tasks can be selected through task-specific evaluation only. Without attempting to be complete in our comparison, the ones described in the rest of the section were selected.

a: BAG OF n-GRAMS (BoN)

To generate vector representations of the text corpus, the bag of n-grams model, with $n \in \{1, 2\}$, was combined with several preprocessing and feature selection methods. For each of the two training datasets, two variations were created, one where the text was used as it is and one where the text was preprocessed. The goal was a comparison of the preprocessing impact on the results, considering the fact that bi-grams are included as features in addition to the single words (uni-grams). The preprocessing pipeline starts with tokenization, followed by stopwords removal and lemmatization, using the NLTK English stopwords corpus and the WordNet Lemmatizer. The features were derived by combining single words and bi-grams appearing in the training dataset, as it has been shown that adding bi-grams can positively influence the performance on certain tasks [51]. The features (n-grams) were weighted using the TF-IDF method, already briefly summarized in Section IV-B, with a required minimum document frequency of the features equal to five documents, in order to discard features that appear very rarely in the corpus. In this particular case, the term "document" refers to one lead section excerpt. This again resulted in high-dimensional, sparse feature vectors $v_i \in \mathbb{R}^{|V|}$, where $|V|$ is the vocabulary size, including both single words and bi-grams. To retain only the most discriminative features, the χ^2 test of dependence between each feature and the target variable was used. The test measures the dependence between two stochastic variables, giving higher scores to the features more related to the target variable and lower scores to those less related. To select the most relevant features, our analysis took into consideration the significance (p-value) of each dependence score, filtering out all insignificant scores, with a p-value higher than 0.1. This resulted in filtering out low scored features, and the number of selected features for each dataset is given in Table 1. The results are consistent with our expectations, and the vector dimensionality was significantly reduced. Additional variations of all four datasets were further produced by retaining only the top k features, $k \in \{50, 100, 200, 400, 800, 1600, 3200\}$, resulting in 32 (slightly) different dataset encodings.

b: PRETRAINED TEXT ENCODERS

To utilize the benefits of the transfer learning concept, several state-of-the-art pre-trained general-purpose models for encoding short text into low-dimensional vector space were utilized. The architecture suggested by Cer *et al.* [50], Universal Sentence Encoder (USE) - Transformer, trained using multitask learning on various unsupervised and supervised tasks and available in TensorFlow Hub,[18] was used as text encoder to produce 512-dimensional embeddings of the text. No preprocessing was applied to the text. Two pretrained InferSent [49] models for text encoding, trained on natural language inference data,[19] version 1 trained using GloVe word embeddings and version 2 trained over FastText word embeddings were used as well, producing output vectors of dimension 4096. Table 1 summarizes the embeddings dimensions on the training datasets.

3) CLASSIFICATION

It has been observed that imbalanced datasets can cause difficulties for the standard classifiers that perform well on balanced data, and each classifier has its specific challenges [52]. The conventional approaches for mitigating this issue involve data-level methods (e.g., undersampling, oversampling), algorithm-level methods (e.g., cost-sensitive learning, one-class learning), or a combination. In this work, three different types of learning algorithms, i.e., Support Vector Machines (SVM) with linear kernel, decision trees, and fully-connected neural network, are trained with the different text encodings of the two training datasets. Different types of classifiers were considered due to the text encoders diversity with regard to the features they extract and their embeddings dimension. The cost-sensitive learning method was incorporated into a large number of different classifier hyperparameter configurations. The SVM classifier was trained using the squared hinge loss function and L2 regularization with different values of the regularization parameter C. The decision tree classifier used the Gini impurity to measure the quality of the split. A varying maximal tree depth d_{max} was evaluated while keeping the minimum number of samples in a node before splitting to a value of two, the minimum number of samples in

[18] https://tfhub.dev/google/universal-sentence-encoder-large/3
[19] https://github.com/facebookresearch/InferSent

TABLE 2. Examples of Wikipedia articles missclassified by the text classifier (false positives), but correctly classified as negative by the node classifier. The articles lead section excerpts either contain common syntactic patterns for disciplines or a sentence defining a discipline studying the concept that the article describes, but not defining the actual concept itself.

Title	Lead Section Excerpt
Conversation	Conversation is interactive, communication between two or more people. The development of conversational skills and etiquette is an important part of socialization. The development of conversational skills in a new language is a frequent focus of language teaching and learning. *Conversation analysis is a branch of sociology which studies the structure and organization of human interaction, with a more specific focus on conversational interaction.*
Hematologic disease	Hematologic diseases are disorders which primarily affect the blood. *Hematology includes the study of these disorders.*
Memorization	Memorization is the process of committing something to memory. Mental process undertaken in order to store in memory for later recall items such as experiences, names, appointments, addresses, telephone numbers, lists, stories, poems, pictures, maps, diagrams, facts, music or other visual, auditory, or tactical information. *The scientific study of memory is part of cognitive neuroscience, an interdisciplinary link between cognitive psychology and neuroscience.*
Cross-sectional study	In medical research and social science, a cross-sectional study (also known as a cross-sectional analysis, transversal study, prevalence study) is a type of observational study that analyzes data collected from a population, or a representative subset, at a specific point in time-that is, cross-sectional data.

leaf nodes to one, and setting no limit of the leaf nodes count. The neural network had a single hidden layer of n ReLU units, $n \in \{50, 100, 200, 400, 800, 1600\}$, as applicable based on the input embeddings dimension. It was trained with the Adam optimization algorithm ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$), using a batch size of 512. Different values of the L2 regularization parameter were evaluated. Because of the combination of several text encoders and diverse types of classifiers, trained models deciding based on different types of features (characteristics of the text) were expected as a result. To utilize their strengths and mitigate their weaknesses, ensemble learning was used to build the final classifier as a combination of several diverse best-performing models, deriving the final prediction using the majority voting technique. Considering the evaluation metrics, the most commonly used metric, i.e., the accuracy, is not appropriate when the imbalance ratio is significant, as in this case. As a result, the precision, recall, and the F1 measure, when the rare class is considered as relevant/positive class and the prevalent class as a negative class, were used for classifiers evaluation.

G. NODE CLASSIFIER

The purpose of the Text Classifier module was the detection of Wikipedia concepts which potentially refer to disciplines by learning the syntactic and semantic patterns that are most common in disciplines lead section excerpts. Nevertheless, not all Wikipedia articles which contain such patterns in the first several sentences refer to a discipline, and Table 2 gives several such examples. Through a combination of the results from the Text Classifier with graph-based analyses, the Node Classifier has the purpose of assigning the final probability that one Wikipedia article refers to a discipline. At this stage, the problem was defined as a binary classification over a balanced dataset.

1) TRAINING/TEST DATASET CREATION

The training and test datasets were created based on the positively classified samples by the Text Classifier on 2017-06-01 Wikipedia export, after their manual evaluation. The positive samples were selected based on a matching of the positively classified article titles with the titles of the disciplines present in CIP. The negative samples were randomly selected from those marked as such in the manual evaluation. A balanced dataset of 500 positive and negative samples, or a total of 1,000 samples, was created and divided randomly into training/test set with ratio 80/20%.

TABLE 3. Graph centrality metrics used as node features.

Graph Type	Node Features
Directed	In degree; Out degree; Closeness centrality (on the regular and reversed graph); Betweenness centrality; Load centrality; Harmonic centrality; Eigenvector centrality; PageRank; Authority & Hub (HITS)
Undirected	Degree; Closeness centrality; Betweenness centrality; Load centrality; Harmonic centrality; Largest maximal clique size; Number of maximal node cliques

2) NODE REPRESENTATION

The samples were represented by a set of features composed of samples probability/confidence scores assigned by the Text Classifier and various graph-based centrality metrics calculated over a graph of all candidate disciplines. The directed graph was constructed using the candidates as graph nodes V, and their articles interlinks in Wikipedia as directed edges. An edge pointing from node (candidate discipline) V_1 to node (candidate discipline) V_2 exists if the Wikipedia article of V_1 contains reference to the article of V_2. The set of features consists of 18 graph-based centralities (Table 3) calculated over the directed graph and its undirected version, and three probability/confidence scores assigned by the Text Classifier.

3) NODE CLASSIFICATION

Different types of classifiers and their hyperparameter combinations were compared on the training dataset using 5-fold cross-validation. The set of evaluated classifiers includes logistic regression, SVM with linear, radial basis function (RBF) and polynomial of degree 3 kernels, decision trees, k-nearest neighbors, and fully-connected neural network with one hidden layer of $n \in \{5, 10, 15, 20\}$ ReLU units. The neural networks were trained using the Adam optimization algorithm. Due to the limited number of training samples, the goal was to select a classifier that exhibits high accuracy and low standard deviation in the cross-validation.

V. RESULTS AND DISCUSSION

A. TEXT CLASSIFIER RESULTS

The text embeddings produced by the different encoder variations based on the BoN model combined with TF-IDF weighting, preprocessing, and feature selection methods, the Universal Sentence Encoder and InferSent pre-trained models, were used as an input in a large number of configurations of the three different types of classifiers, described in Section IV-F.3. Two different training datasets, one consisting of samples from Wikipedia solely and one combining samples from both Wikipedia and CIP, were encoded with all text encoders and used to train the classifiers, evaluated on a test dataset consisting only of Wikipedia samples. This section presents a summary of the results, without implying any generality with other types of tasks from this or another domain. The evaluation measurements presented here and used to compare the trained models are the precision, recall, and F1 on the positive (rare) class, achieved on the test dataset.

In line with the previously stated expectations, the different text encoders achieved their best results with different types of classifiers. Fig. 4 gives the maximum F1, which the text encoders scored with the different classifiers when trained on each of the two different training datasets. A maximum F1 of 0.91 was achieved by the InferSent encoder working with FastText word embeddings, in combination with SVM with a linear kernel. From the presented results and the findings in the relevant related work, it has become evident that different types of encoders capture different characteristics of the texts in the fixed-length vectors, and it can be useful to combine them for improved results. For each of the different types of encoders, BoN, USE Transformer, and InferSent, one best-performing model was selected, the results of which are summarized in Table 4. The three models predictions were combined through majority voting, achieving F1 of 0.96 on the test dataset. The final architecture of the Text Classifier is illustrated in Fig. 5.

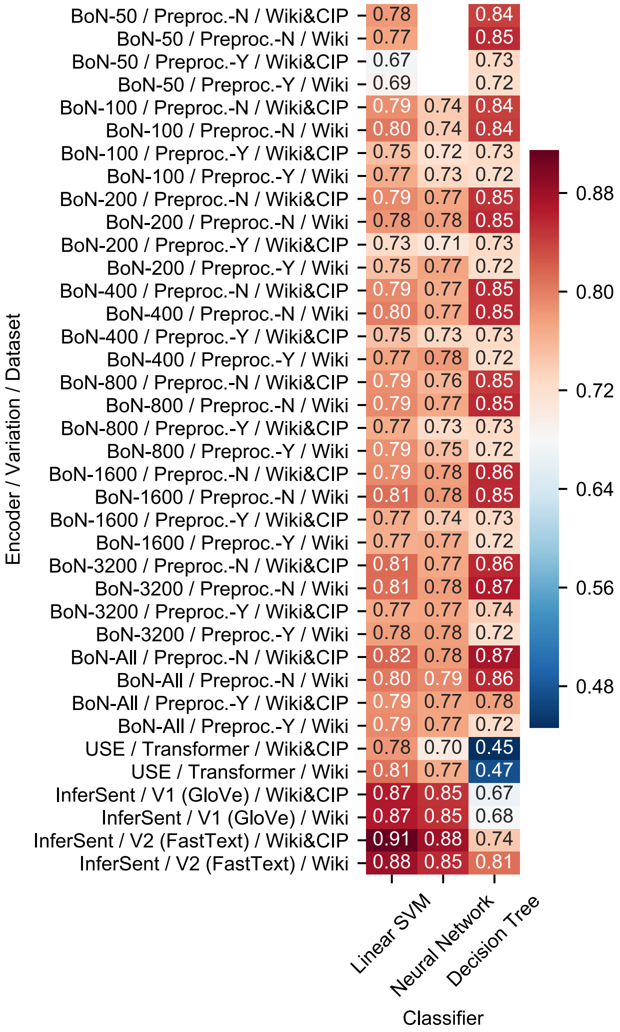

FIGURE 4. Max F1 on the test dataset of the text encoders and various configurations of classifiers trained on the two training dataset. The vertical axis states the text encoder and the training dataset which the different classifier configurations were trained on.

The results of the three selected models and majority voting classifier (for details, please refer to the project GitHub repository) show that the false negative (FN) samples are not among the well-established, top-level disciplines, but sub-disciplines which are rather ambiguous for humans as well. Furthermore, the false positive (FP) samples are concepts closely related to science, which contain in their lead section excerpts some of the most important terms used in defining disciplines (e.g., *science*, *study*, *discipline*), or a definition of a discipline studying them. Some of the miss-classified samples (e.g., *Dromography*) can even be considered as wrongly classified during the test dataset creation. Nevertheless, the majority voting classifier manages to correct most of the errors by the individual models, resulting in only four miss-classified samples on the test dataset. However, it is not able to address the miss-classified samples due to human factors such as lead section excerpts containing a definition of discipline studying the concept which the Wikipedia article refers to, task of the subsequent module, Node Classifier.

The results further indicate that on this specific task, almost all encoders score comparably well when combined with an appropriate type of classifier. The TF-IDF weighting of the uni-grams and bi-grams appearing in the text, combined with several pre-processing steps, appears to be an equally good text encoding method on this specific task as the more complex pre-trained encoders. When followed by feature selection techniques, the dimensionality of the vectors can be significantly reduced to a set of most distinctive features. The discipline detection task proves to be a task where a small set of words and phrases carry the most of the distinctive power, and even encoders retaining only the 50 top-scoring features score comparably well (F1 of 0.85) as those retaining several thousand features.

TABLE 4. Precision (P), Recall (R) and F1 of the three selected text classification models and the majority voting classifier achieved on the rare (positive) class. The values of the varying hyperparameters are given next to the classifier name. The rest of the hyperparameter values are given in Section IV-F.3.

Model	P	R	F1
BoN-All (Preproc.-No) & Decision Tree ($d_{max} = 10$)	0.85	0.90	0.87
USE (Transformer) & Linear SVM ($C = 10$)	0.86	0.76	0.81
InferSent (FastText) & Linear SVM ($C = 1$)	0.87	0.96	0.91
Majority Voting	0.94	0.98	0.96

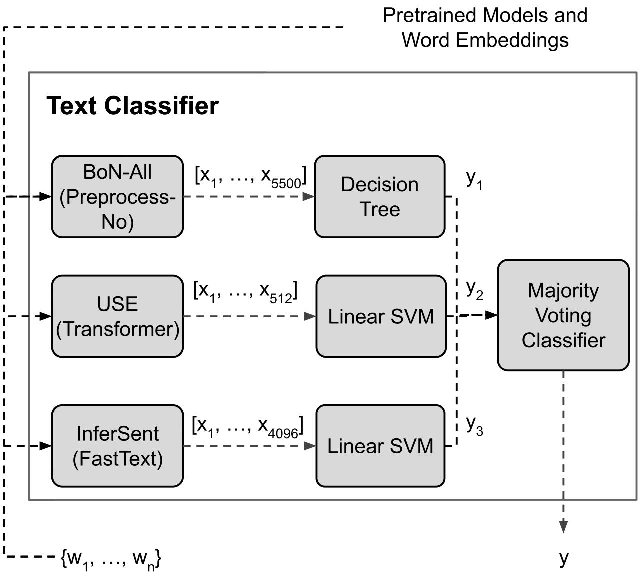

FIGURE 5. Text Classifier architecture.

B. NODE CLASSIFIER RESULTS

The classifiers mean accuracy and standard deviation from the 5-fold cross-validation were compared, and the classifier configuration with the highest accuracy and low standard deviation was selected. It is a fully-connected neural network with one hidden layer of 5 ReLU units, trained using the Adam optimization algorithm with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\alpha = 0.001$, a batch size of 64 and L2 regularization of 0.01. The configuration achieved an average accuracy of 0.883(±0.02) in the cross-validation, and the final model achieved 0.885 accuracy on the test dataset at a probability threshold of 0.5. Our manual analyses of the misclassified samples led us to a conclusion that many of them are rather ambiguous for human evaluators as well, particularly the false positive samples.

C. EVALUATION ON DIFFERENT WIKIPEDIA EXPORTS

Further evaluation of the methodology was done by processing Wikipedia exports from four different years and a comparison of the results. The exports are from 2015-06-02, 2016-06-01, 2017-06-01 (also used for training purposes) and 2018-11-20. This section provides statistical and visual insights into the results.

TABLE 5. Number of retained articles after different processing stages, by Wikipedia export. The score threshold is 0.67 for the text classifier (vote by at least two classifiers out of three) and 0.5 for the node classifier.

Wikipedia Export	Text Extractor	Basic Filter	Text Classifier	Node Classifier
2015-06-02	4,356,236	1,308,681	4,936	2,337
2016-06-01	4,620,844	1,378,199	5,048	2,383
2017-06-01	4,851,262	1,439,207	5,123	2,445
2018-11-20	5,152,132	1,520,711	5,255	2,501

TABLE 6. Performance of the node classifier trained on the export from 2017-06-01 and evaluated on the other exports. The test datasets of the other exports contain the same disciplines as the test dataset of 2017-06-01 export, but with feature values calculated on the appropriate export.

Wikipedia Export	Test Dataset Size	Accuracy
2015-06-02	173	0.879
2016-06-01	185	0.881
2017-06-01	200	0.885
2018-11-20	184	0.886

Table 5 provides a summary of the total number of articles retained after the initial text/metadata extraction, the number of articles retained after the basic filtering, the total number of detected candidate disciplines and the actual detected disciplines per year, at the specified thresholds. To evaluate if the performance of the discipline detection model trained on the Wikipedia export from 2017-06-01 decreases on the other exports, the test dataset was mapped to the data from the other three exports and evaluated. Table 6 summarizes the results, indicating that a significant decrease in the accuracy is not present, although it tends to slightly decrease for the past exports, as the time difference increases. Further analyses of the period in which one model is applicable are planned as part of the future work when the observed period would be extended and the test dataset enlarged. Fig. 6 provides a brief overview of the detected disciplines by export, where the detected disciplines and their interlinks in Wikipedia are visualized. The directed graph is created in the same manner as described in Section IV-G.2. For a better overview, the nodes are grouped in communities using the greedy modularity maximization algorithm [53], as implemented in NetworkX. The node labels of the 100 disciplines with the highest probability score are displayed, while the size of the nodes resembles their probability score and their color the community they belong to. At this stage, we do not go into in-depth community structure analyses, since the purpose of introducing the communities was the clarity of the visualizations solely. The visualizations were created with Gephi [54]. For details on the 100 labeled disciplines in the visualizations, please refer to the project GitHub repository.

The overall conclusion from the evaluation is that the results are satisfactory to a large extent and, at the same time, very useful in the identification of the aspects which can be further improved to reach even higher discipline detection accuracy. One such improvement is the lead section

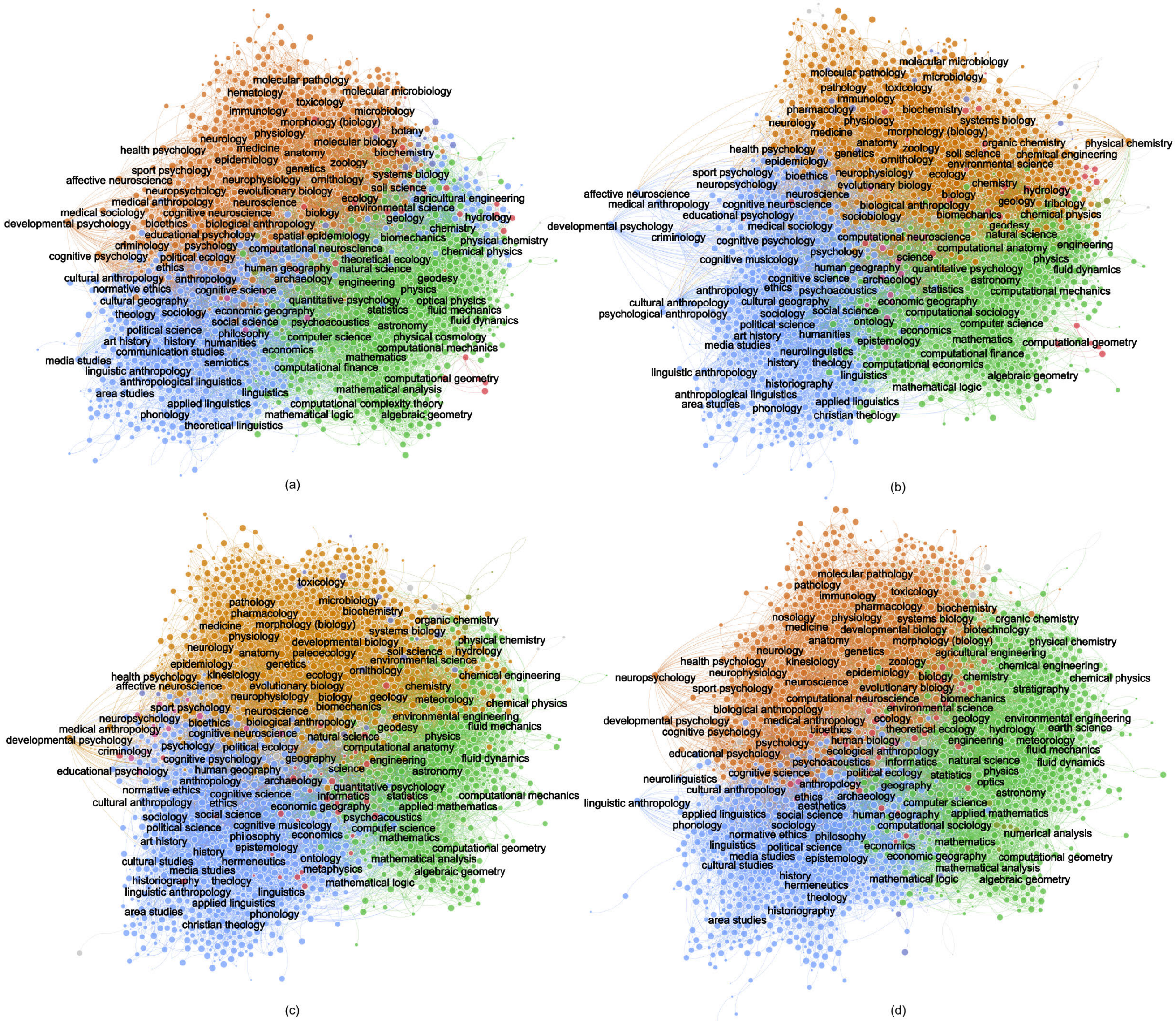

FIGURE 6. **Detected disciplines graph by export. (a) 2015-06-02. (b) 2016-06-01. (c) 2017-06-01. (d) 2018-11-20.**

excerpts extraction logic to distinguish better the lead section sentences that define the article's topic from the sentences defining other related topics that should be left out. Altogether, the presented methodology is developed based on extensive domain-specific analyses, in circumstances of no clear definition of the concept of ''academic discipline'' and non-existence of a sufficiently large dataset that can be used for data-driven methodology development and proper evaluation. The implemented software application is capable of working with large data volumes, offering an opportunity to restore historical states of the domain as long as Wikipedia's historical exports are available, and opening a plentitude of possibilities for subsequent data analyses, as well as complementing the methodologies based on scientific publication analyses.

VI. CONCLUSION

Wikipedia represents a fast-changing resource with a collaborative editing policy, which ensures that it contains current information at any point in time, reliable to a large extent. Examining the opinion of the world-wide community on what does and what does not represent an academic discipline through analyses of Wikipedia may be very informative, considering the fact that a clear and definite ''ground truth'' does not exist, and no approach can be fully evaluated. This work analyzes Wikipedia's usefulness in determining the disciplines that exist at a certain moment through a data-driven methodology for detection of the concepts which Wikipedia's community has defined as disciplines at that same moment. Evaluated on four different Wikipedia exports, the methodology achieves high accuracy in the selected evaluation context.

The future work will compare the achieved results with larger number of expert classification schemes in order to further refine the methodology, evaluate challenges like concept drift to detect if and in what time interval the trained models should be updated, study the stability of the results across exports and make the results available in a user-friendly manner through a web application.

REFERENCES

[1] M. Gibbons, *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Newbury Park, CA, USA: Sage, 1994.

[2] R. Stichweh, "The sociology of scientific disciplines: On the genesis and stability of the disciplinary structure of modern science," *Sci. Context*, vol. 5, no. 1, pp. 3–15, 1992.

[3] R. Stichweh, "Scientific disciplines, history of," in *International Encyclopedia of the Social & Behavioral Sciences*. Dec. 2001, pp. 13727–13731. [Online]. Available: https://www.researchgate.net/publication/277686080_Scientific_Disciplines_History_of, doi: 10.1016/B0-08-043076-7/03187-9.

[4] J. Surowiecki, "The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business," *Economies, Societies Nations*, vol. 296, 2004.

[5] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie," *World Wide Web*, vol. 1, no. 2, p. 19, 2007.

[6] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, pp. 900–901, Dec. 2005. [Online]. Available: https://www.nature.com/articles/438900a

[7] M. S. Rajagopalan, V. K. Khanna, Y. Leiter, M. Stott, T. N. Showalter, A. P. Dicker, and Y. R. Lawrence, "Patient–oriented cancer information on the Internet: A comparison of Wikipedia and a professionally maintained database," *J. Oncol. Pract.*, vol. 7, no. 5, pp. 319–323, Sep. 2011.

[8] J. Kräenbring, T. Monzon Penza, J. Gutmann, S. Muehlich, O. Zolk, L. Wojnowski, R. Maas, S. Engelhardt, and A. Sarikas, "Accuracy and completeness of drug information in Wikipedia: A comparison with standard textbooks of pharmacology," *PLoS ONE*, vol. 9, no. 9, Sep. 2014, Art. no. e106930.

[9] F. Viegas, M. Wattenberg, J. Kriss, and F. Ham, "Talk before you type: Coordination in Wikipedia," in *Proc. 40th Annu. Hawaii Int. Conf. Syst. Sci. (HICSS)*, 2007, p. 78.

[10] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl, "A jury of your peers: Quality, experience and ownership in Wikipedia," in *Proc. 5th Int. Symp. Wikis Open Collaboration (WikiSym)*, 2009, p. 15.

[11] I. Kubiszewski, T. Noordewier, and R. Costanza, "Perceived credibility of Internet encyclopedias," *Comput. Educ.*, vol. 56, no. 3, pp. 659–667, Apr. 2011.

[12] N. Jullien. (2012). *What We Know About Wikipedia: A Review of the Literature Analyzing the Project(s)*. [Online]. Available: https://ssrn.com/abstract=2053597

[13] (2019). *Introduction to the Classification of Instructional Programs: 2020 Edition (CIP-2020)*. [Online]. Available: https://nces.ed.gov/ipeds/cipcode/Files/2020_CIP_Introduction.pdf

[14] (2007). *Revised Field of Science and Technology (FOS) Classification in the Frascati Manual*. [Online]. Available: http://www.oecd.org/science/inno/38235147.pdf

[15] (2019). *Introduction to the Dewey Decimal Classification*. [Online]. Available: https://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf

[16] A. Slavic, "UsUse of the universal decimal classification: A world-wide survey," *J. Documentation*, vol. 64, no. 2, pp. 211–228, Mar. 2008.

[17] UDC Consortium. *UDC Structure & Tables*. Accessed: Sep. 15, 2019. [Online]. Available: http://www.udcc.org/index.php/site/page?view=about_structure

[18] L. M. Chan, S. S. Intner, and J. Weihs, *Guide to the Library of Congress Classification*. Santa Barbara, CA, USA: ABC-CLIO, 2016.

[19] L. Waltman and N. J. Van Eck, "A new methodology for constructing a publication-level classification system of science," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 12, pp. 2378–2392, Dec. 2012.

[20] Q. Wang, "A bibliometric model for identifying emerging research topics," *J. Assoc. Inf. Sci. Technol.*, vol. 69, no. 2, pp. 290–304, Feb. 2018.

[21] H. Small, K. W. Boyack, and R. Klavans, "Identifying emerging topics in science and technology," *Res. Policy*, vol. 43, no. 8, pp. 1450–1467, Oct. 2014.

[22] K. W. Boyack and R. Klavans, "Creation of a highly detailed, dynamic, global model and map of science," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 670–685, Apr. 2014.

[23] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, and E. Motta, "The computer science ontology: A large-scale taxonomy of research areas," in *The Semantic Web—ISWC 2018*, D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, and E. Simperl, Eds. Cham, Switzerland: Springer, 2018, pp. 187–205. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-00668-6_12

[24] F. Osborne and E. Motta, "Klink-2: Integrating multiple Web sources to generate semantic topic networks," in *The Semantic Web—ISWC 2015*, M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, K. Thirunarayan, and S. Staab, Eds. Cham, Switzerland: Springer, 2015, pp. 408–424.

[25] A. A. Salatino, F. Osborne, and E. Motta, "How are topics born? Understanding the research dynamics preceding the emergence of new areas," *PeerJ Comput. Sci.*, vol. 3, p. e119, Jun. 2017.

[26] A. Suominen and H. Toivanen, "Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 10, pp. 2464–2476, 2016.

[27] L. Leydesdorff and I. Rafols, "A global map of science based on the ISI subject categories," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 2, pp. 348–362, 2009.

[28] L. Leydesdorff, S. Carley, and I. Rafols, "Global maps of science based on the new web-of-science categories," *Scientometrics*, vol. 94, no. 2, pp. 589–593, 2013.

[29] X. Meng, X. Liu, Y. Tong, W. Glänzel, and S. Tan, "Multi-view clustering with exemplars for scientific mapping," *Scientometrics*, vol. 105, no. 3, pp. 1527–1552, 2015.

[30] J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva, "Clickstream data yields high-resolution maps of science," *PLoS ONE*, vol. 4, no. 3, p. e4803, 2009.

[31] D. Chavalarias and J.-P. Cointet, "Phylomemetic patterns in science evolution—The rise and fall of scientific fields," *PLoS ONE*, vol. 8, no. 2, Feb. 2013, Art. no. e54847.

[32] M. Herrera, D. C. Roberts, and N. Gulbahce, "Mapping the evolution of scientific fields," *PLoS ONE*, vol. 5, no. 5, 2010, Art. no. e10355.

[33] A. A. Salah, C. Gao, K. Suchecki, and A. Scharnhorst, "Need to categorize: A comparative look at the categories of universal decimal classification system and wikipedia," *Leonardo*, vol. 45, no. 1, pp. 84–85, 2012.

[34] J. Minguillón, M. Lerga, E. Aibar, J. Lladós-Masllorens, and A. Meseguer-Artola, "Semi-automatic generation of a corpus of Wikipedia articles on science and technology," *El Prof. Inform.*, vol. 26, no. 5, pp. 995–1004, 2017.

[35] A. Joorabchi and A. E. Mahdi, "Towards linking libraries and Wikipedia: Automatic subject indexing of library records with Wikipedia concepts," *J. Inf. Sci.*, vol. 40, no. 2, pp. 211–221, 2014.

[36] A. Joorabchi and A. E. Mahdi, "Improving the visibility of library resources via mapping library subject headings to Wikipedia articles," *Library Hi Tech*, vol. 36, no. 1, pp. 57–74, 2018.

[37] J. Yoon, J. Yun, and W.-S. Jung, "Build up of a subject classification system from collective intelligence," 2018, *arXiv:1804.00026*. [Online]. Available: https://arxiv.org/abs/1804.00026

[38] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[39] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2018.

[40] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Newton, MA, USA: O'Reilly Media, 2009.

[41] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*, May 2010, pp. 45–50. [Online]. Available: http://is.muni.cz/publication/884893/en

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[43] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python Sci. Conf.*, G. Varoquaux, T. Vaught, and J. Millman, Eds. Pasadena, CA, USA, 2008, pp. 11–15. [Online]. Available: http://conference.scipy.org/proceedings/SciPy2008/paper_2/

[44] Z. S. Harris, "Distributional structure," *Word*, vol. 10, nos. 2–3, pp. 146–162, 1954.

[45] A. Dimitrovski, A. Gjorgjevikj, and D. Trajanov, "Courses content classification based on Wikipedia and CIP taxonomy," in *ICT Innovations 2017*, D. Trajanov and V. Bakeva, Eds. Cham, Switzerland: Springer, 2017, pp. 140–153. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-67597-8_14

[46] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[47] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3294–3302.

[48] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," 2016, *arXiv:1602.03483*. [Online]. Available: https://arxiv.org/abs/1602.03483

[49] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," 2017, *arXiv:1705.02364*. [Online]. Available: https://arxiv.org/abs/1705.02364

[50] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, and C. Tar, "Universal sentence encoder," 2018, *arXiv:1803.11175*. [Online]. Available: https://arxiv.org/abs/1803.11175

[51] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2012, pp. 90–94.

[52] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.

[53] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, 2004, Art. no. 066111.

[54] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *Proc. 3rd Int. AAAI Conf. Weblogs Social Media*, 2009.

ANA GJORGJEVIKJ received the bachelor's degree in computer science and engineering and the master's degree in computer networks and e-technologies from the Ss. Cyril and Methodius University in Skopje, in 2010 and 2014, respectively, where she is currently pursuing the Ph.D. degree in computer science with a particular focus on deep learning and natural language processing. She has been a Software Engineer, since 2010. Her research interests include data science, machine learning, natural language processing, and knowledge representation.

KOSTADIN MISHEV received the bachelor's degree in informatics and computer engineering and the master's degree in computer networks and e-technologies from the Ss. Cyril and Methodius University in Skopje, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Engineering. He is also a Teaching and a Research Assistant with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University of Skopje. His research interests include data science, semantic web, natural language processing, enterprise application architectures, web technologies, and computer networks.

DIMITAR TRAJANOV received the Ph.D. degree. He is currently a Professor and the Head of the Department of Information Systems and Network Technologies, Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. He is also the Leader of the Regional Social Innovation Hub established in 2013 as cooperation between UNDP and the Faculty of Computer Science and Engineering. He is the author of more than 150 journals and conference papers and seven books. He has been involved in more than 50 research and industry projects, of which in more than 30 projects as a Project Leader. His research interests include data science, machine learning, NLP, FinTech, semantic Web, open data, sharing economy, social innovation, e-commerce, entrepreneurship, technology for development, mobile development, and climate change.