Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Citation:
Sawicki, Jan; Ganzha, Maria; Paprzycki, Marcin. The state of the art of Natural Language Processing – a systematic automated review of NLP literature using NLP techniques. Data Intelligence. DOI: 10.1162/dint_a_00213

The state of the art of Natural Language Processing – a systematic automated review of NLP literature using NLP techniques

Jan Sawicki[†], Maria Ganzha[†], Marcin Paprzycki[†]

Abstract

Nowadays, natural language processing (NLP) is one of the most popular areas of, broadly understood, artificial intelligence. Therefore, every day, new research contributions are posted, for instance, to the arXiv repository. Hence, it is rather difficult to capture the current "state of the field" and thus, to enter it. This brought the idea of applying state-of-the-art NLP techniques to analyse the NLP-focused literature. As a result, (1) meta-level knowledge, concerning the current state of NLP has been captured, and (2) a guide to use of basic NLP tools is provided. It should be noted that all the tools and the dataset described in this contribution are publicly available. Furthermore, the originality of this review lies in its full automation. This allows easy reproducibility and continuation and updating of this research in the future as new researches emerge in the field of NLP.

Keywords: natural language processing, text processing, literature survey, keyword search, keyphrase search, text embeddings, text summarizations

1. Introduction

Natural language processing (NLP) is rapidly growing in popularity in a variety of domains, from closely related, like semantics [1, 3] and linguistics [2, 4] (e.g. inflection [176], phonetics and onomastics [175], automatic text correction [177]), named entity recognition [179, 178] to distant ones, like biobliometry [6], cybersecurity [7], quantum mechanics [8, 9], gender studies [10, 5], chemistry [11] or orthodontia [12]. This, among others, brings an opportunity, for early-stage researchers, to enter the area. Since NLP can be applied to many domains and languages, and involves use of many techniques and approaches, it is important to realize where to start.

This contribution attempts at addressing this issue, by applying NLP techniques to analysis of NLP-focused literature. As a result, with a fully automated, systematic, visualization-driven literature analysis, a guide to the state-of-the-art of natural language processing is presented. In this way, two goals are achieved. (1) Providing introduction to NLP for scientists entering the field, and (2) supporting possible knowledge update for experienced researchers. The main research questions (RQs) considered in this work are:

[†]Corresponding author; Warsaw University of Technology; email: jan.sawicki2.dokt@pw.edu.pl; ORCID: 0000-0002-8930-7564)

[†]Warsaw University of Technology; email: maria.ganzha@pw.edu.pl; ORCID: 0000-0001-7714-4844) [†]Polish Academy of Sciences; email: paprzyck@ibspan.waw.pl; ORCID: 0000-0002-8069-2152)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

RQ1: What datasets are considered to be most useful?

RQ2: Which languages, other than English, appear in NLP research?

RQ3: What are the most popular fields and topics in current NLP research?

RQ4: What particular tasks and problems are most often studied?

RQ5: Is the field "homogenous", or are there easily identifiable "subgroups"?

RQ6: How difficult is it to comprehend the NLP literature?

Taking into account that the proposed approach is, itself, anchored in NLP, this work is also an illustration of how selected standard NLP techniques can be used in practice, and which of them should be used for which purpose. However, it should be made clear that considerations presented in what follows should be treated as "illustrative examples", not "strict guidelines". Moreover, it should be stressed that none of the applied techniques has been optimized to the task (e.g. no hyperparameter tuning has been applied). This is a deliberate choice, as the goal is to provide an overview and "general ideas", rather than overwhelm the reader with technical details of individual NLP approaches. For technical details, concerning optimization of mentioned approaches, reader should consult referenced literature.

The whole analysis has been performed in Python – a programming language which is ubiquitous in data science research and projects for years [17, 16, 21, 18, 20, 19]. Python was also chosen for the following reasons:

- It provides a heterogeneous environment
- It allows use of Jupyter Notebooks[1], which allow quick and easy prototyping, testing and code sharing
- There exists an abundance of data science libraries[2], which allow everything from acquiring the dataset, to visualizing the result
- It offers readability and speed in development [171]

Presented analysis follows the order of research questions. To make the text more readable, readers are introduced to pertinent NLP methods in the context of answering individual questions.

2. Data and preprocessing

At the beginning of NLP research, there is always data. This section introduces the dataset consisting of research papers used in this work, and describes how it was preprocessed.

[1] `https://jupyter.org`
[2] `https://pypi.org`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

2.1. Data used in the research

To adequately represent the domain, and to apply NLP techniques, it is necessary to select an abundant, and well-documented, repository of related texts (stored in a digital format). Moreover, to automatize the conducted analysis, and to allow easy reproduction, it is crucial to choose a set of papers, which can be easily accessed, e.g. a database with a functional Application Programming Interface (API). Finally, for obvious reasons, open access datasets are the natural targets for NLP-oriented work.

In the context of this work, while there are multiple repositories, which contain NLP-related literature, the best choice turned out to be arXiv (for the papers themselves, and for the metadata it provided), combined with the Semantic Scholar (for the "citation network" and other important metadata; see Section 3.3.1).

Note that other datasets have been considered, but were not selected. Reasons for this decision have been summarized in Table 1.

Table 1: Consideration regarding databases not used in the analysis

Database	The reason for in applicability in this research task
Google Scholar	Google Scholar does not contain actual data (text, PDF, etc.) of any work – there are only links to other databases. Moreover, performed tests determined that the API (Python "scholarly" library) works well with small queries, but fetching information about thousands of papers results in download rate limits, and temporary IP address blocking. Finally, Google Scholar is criticized, among others, for excessive secrecy [14], biased search algorithms [13], and incorrect citation counts [15].
PubMed	PubMed is mainly focused on medical and biological papers. Therefore, the number of works related to NLP is somewhat limited, and difficult to identify using straightforward approaches.
ResearchGate	There are two main problems with ResearchGate, as seen from the perspective of this work: lack of easy-accessible API and restrictions on some articles' availability (large number of papers has to be requested from authors – and such requests may not be fulfilled, or wait time may be excessive).
Scopus	The Scopus API is not fully open-access, and has restrictions on the number of requests that can be issues within a specific time.
JSTOR	Even though the JSTOR website [3] declares that API exists, the link does not provide any information about it (404 not found).
Microsoft Academic	The Microsoft Academic API is very well documented, but it does not provide true open access (requires a subscription key). Moreover, it does not contain the actual text of works; mostly metadata.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

2.1.1. Dataset downloading and filtering

The papers were fetched from arXiv on 26 August 2021. The resulting dataset includes all articles, which have been extracted as a result of issuing the query "natural language processing"[4]. As a result, 4712 articles were retrieved. Two articles were discarded because their PDFs were too complicated for the tools that were used for the text extraction (1710.10229v1 – problems with chart on page 15; 1803.07136v1 – problems with chart on page 6; see, also, section 2.2). Even though the query was *not* bounded by the "time when the article was uploaded to arXiv" parameter, it turned out that a solid majority of the articles had submission dates from the last decade. Specifically, the distribution was as follows:

- 192 records uploaded before 2010-01-01
- 243 records from between (including) 2010-01-01 and 2014-12-31
- 697 records from between (including) 2015-01-01 and 2017-12-31
- 3580 records uploaded after 2018-01-01

On the basis of this distribution, it was decided that there is no reason to impose time constraints, because the "old" works should not be able to "overshadow" the "newest" literature. Moreover, it was decided that it is worth keeping all available publications, as they might result in additional findings (e.g., as what concerns the most original work, described in Section 3.7.4).

Finally, all articles not written in English were discarded, reducing the total count to 4576 texts. This decision, while somewhat controversial, was made to be able to understand the results (by the authors of this contribution) and to avoid complex issues related to text translation. However, it is easy to observe that the number of texts not written in English (and stored in arXiv) was relatively small (< 5%). Nevertheless, this leaves open a question: what is the relationship between NLP-related work that is written in English and that written in other languages. However, addressing this topic is out of scope of this contribution.

2.2. Text preprocessing

Obviously, the key information about a research contribution is contained in its text. Therefore, subsequent analysis applied NLP techniques to texts of downloaded papers. To do this, the following preprocessing has been applied. The PDFs have been converted to plain text, using pdfminer.six (a Python library[5]). Here, notice that there are several other libraries that can also be used to convert PDF to text. Specifically, the following libraries have been tried: pdfminer[6], pdftotree[7], BeautifulSoup[8]. On the basis of performed tests, pdfminer.six was selected, because it provided the simplest API, produced results, which did not have to be further converted (as opposite to, e.g., BeautifulSoup), and performed the fastest conversion.

[4]Specifically, the query had the form `http://export.arxiv.org/api/query?search\_query=all:\%22natural\%20language\%20processing\%22&start=0\&max\_results=10000`. Since such query may take long time to load; to reduce time, one can change the value of the max_results parameter to a smaller number, e.g. 5

[5]`https://pdfminersix.readthedocs.io/en`

[6]`https://github.com/euske/pdfminer`

[7]`https://github.com/HazyResearch/pdftotree`

[8]`https://www.crummy.com/software/BeautifulSoup`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Use of different text analysis methods may require different preprocessing. Some methods, like keyphrase search, work best when the text is "thoroughly cleaned"; i.e. almost reduced to a "bag of words" [167]. This means that, for instance, words are lemmatized, there is no punctuation, etc. However, some more recent techniques (like text embeddings [168]) can (and should) be trained on a "dirty" text, like Wikipedia [169] dumps[9] or Common Crawl[10]. Hence, it is necessary to distinguish between (at least) two levels of text cleaning: (A) "delicately cleaned" text (in what follows, called "Stage 1" cleaning), where only parts insignificant to the NLP analysis are removed, and (B) a "very strictly cleaned" text (called "Stage 2" cleaning). Specifically, "Stage 1" cleaning includes removal of:

- charts and diagrams improperly converted to text,
- arXiv "watermarks",
- references section (which were not needed, since metadata from Semantic Scholar was used),
- links, formulas, misconverted characters (e.g. "ff").

Stage 2 cleaning is applied to the results of Stage 1 cleaning, and consists of the following operations:

- All punctuation, numbers and other non-letter characters were removed, leaving only letters.
- Adposition, adverb, conjunction, coordinating conjunction, determiner, interjection, numeral, particle, pronoun, punctuation, subordinating conjunction, symbol, end of line, space were removed. Parts of speech left after filtering were: verbs, nouns, auxiliaries and "other". The "other" category is usually tagged for meaningless text, e.g. "asdfgh". However, these were not deleted in case the algorithm detected something that was, in fact, important, e.g. domain-specific shortcuts and abbreviations like CNN, RNN, etc.
- Words have been lemmatized.

Note that while individual NLP techniques may require more specific data cleaning, the two (Stage 1 and Stage 2) workflows are generic enough to be successfully applied in the majority of typical NLP applications.

3. Performed experiments, applied methods and analysis of results

This section traverses research questions RQ1 to RQ6 and summarizes the findings for each one of them. Furthermore, it introduces specific NLP methods used to address each question. Interested readers are invited to study referenced literature to find additional details.

3.1. RQ1: finding most popular datasets used in NLP

As noted, a fundamental aspect for all data science projects is the data. Hence, this section summarizes the most popular (open) datasets that are used in NLP research. Here, the information about these datasets (names of datasets) was extracted from the analyzed texts, using Named Entity Recognition and Keyphrase search. Let us briefly summarize these two methods.

[9] `https://dumps.wikimedia.org`
[10] `https://commoncrawl.org`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

3.1.1. Named Entity Recognition – NER

Named Entity Recognition (NER) can be seen as finding an answer to "the problem of locating and categorizing important nouns, and proper nouns, in a text" [180]. Here, automatic methods should facilitate extraction of, among others, named topics, issues, problems, and other "things" mentioned in texts (e.g. in articles). Hence, the spaCy [100] NER model "en-core-web-lg"[11] has been used to extract named entities. These entities have been linked by co-occurrence, and visualized as networks (further described in section 3.4).

SpaCy has been chosen over other models (e.g. transformers [101] pipeline[12]), because it was simpler to use, and performed faster.

3.1.2. Keyphrase search

Another simple and effective way of extracting information from text, is keyword and/or keyphrase search [181, 166]. This technique can be used not only in the preliminary exploratory data analysis (EDA), but also to extract actual and useful findings. Furthermore, keyphrase search is also complementary to, and extends, results of Named Entity Recognition (NER) (Section 3.1.1).

To apply keyphrase search, first, texts were cleaned with Stage 2 cleaning (see Section 2.2). Second, they were converted to phrases (n-grams) of lengths 1-4. Next, two exhaustive lists were created, based on all phrases (n-grams): (a) allowed phrases (609 terms), and (b) banned phrases (1235 terms). The allowed phrases contained word and phrases, which were meaningful for natural language processing or were specific enough to be considered separate, e.g. TF-IDF, accuracy, annotation, NER, taxonomy. The list of banned phrases contains words and phrases, which on their own carried no significant meaning for this research, e.g. bad, big, bit, long, power, index, default. The banned phrases also contained some incoherent phrases, which slipped through the previous cleaning phases. These lists were used to filter the phrases found in the texts. Obtained results were converted to networks of phrase co-occurrence, to visualize phrase importance, and relations between phrases.

3.1.3. Approaches to finding names of most popular NLP datasets

Keyword search was used to extract names of NLP datasets used in collected papers. To properly factor out dataset names and omit noise words, two approaches were applied: unsupervised and list-based.

Unsupervised approach included extracting words (proper nouns detected with Python spaCy[13] library) in the near neighborhood (max 3 words before or after) of words "data", "dataset" and similar.

In list-based approaches, the algorithm looked for particular dataset names that were identified in the three big aggregated lists of NLP datasets[14][15][16].

[11] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.2.0
[12] https://huggingface.co/transformers/main_classes/pipelines.html\#tokenclassificationpipeline
[13] https://spacy.io
[14] https://metatext.io/datasets
[15] https://github.com/niderhoff/nlp-datasets
[16] https://github.com/karthikncode/nlp-datasets

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

3.1.4. Findings related to RQ1; what are the most popular NLP datasets

This section presents the findings, which answer RQ1, i.e. which datasets are most often used in NLP research. To best show datasets that are popular, and outline which are used together, a heatmap has been created. It is presented in Figure 1. In general, a heatmap allows getting not only a general ranking of features (looking only at the diagonal), but also provides the information of correlation of features, or lack thereof.

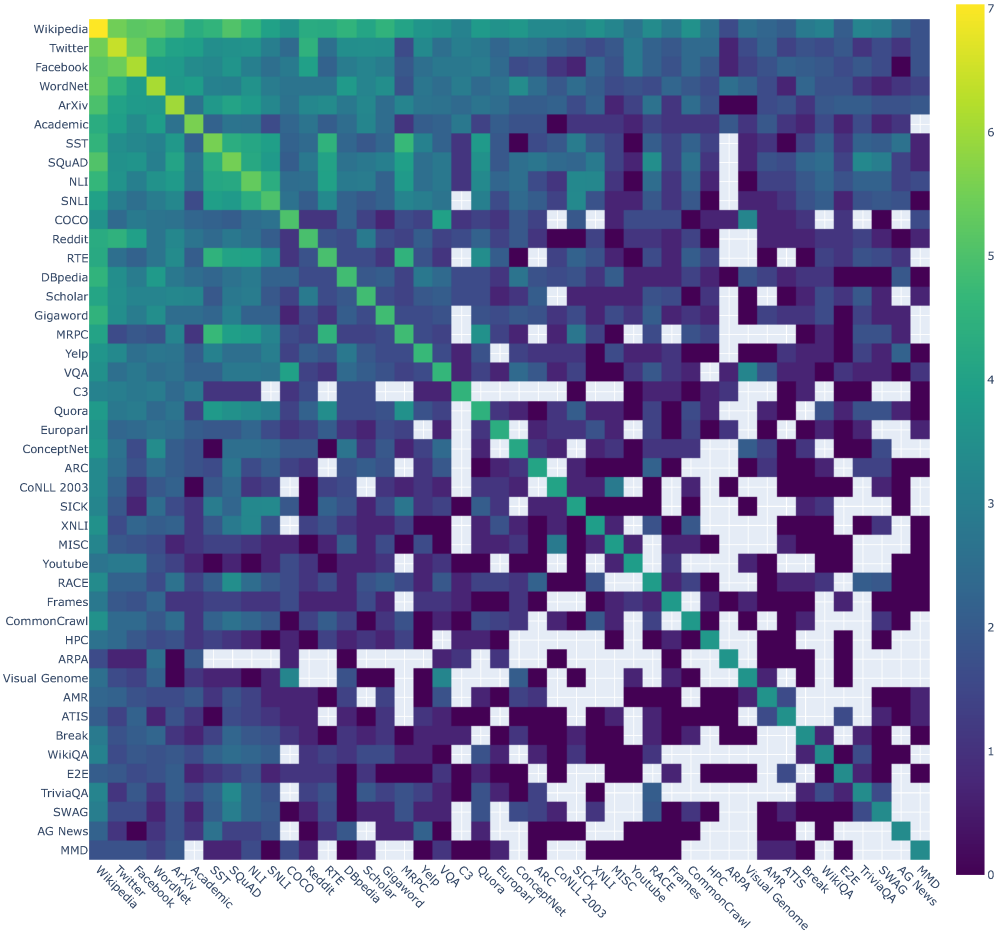

Figure 1: Heatmap of top 10 percentile of NLP datasets co-usage (logarithmic scale).

It can be easily seen that the most popular dataset, used in NLP, is Wikipedia. Among the top 4 most popular datasets, one can find also: Twitter, Facebook, and WordNet. There is a high correlation between use of datasets, which were extracted from Twitter and Facebook, which are very frequently used together. This is both intuitive and observable in articles dedicated to social network analysis [114], social text sentiment analysis[118], social media mining [116] and other

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

social science related texts [117]. Manual checking determined also that Twitter is extremely popular in sentiment analysis and other emotion-related explorations [115].

3.2. Findings related to RQ2: what languages are studied in NLP research

The second research question concerned languages that were analyzed in reported research (not the language the paper was written in). This information was mined using the same two methods, i.e. keyphrase search and NER. The results were represented in two ways. The basic method was a co-occurrence heatmap presented in Figure 2.

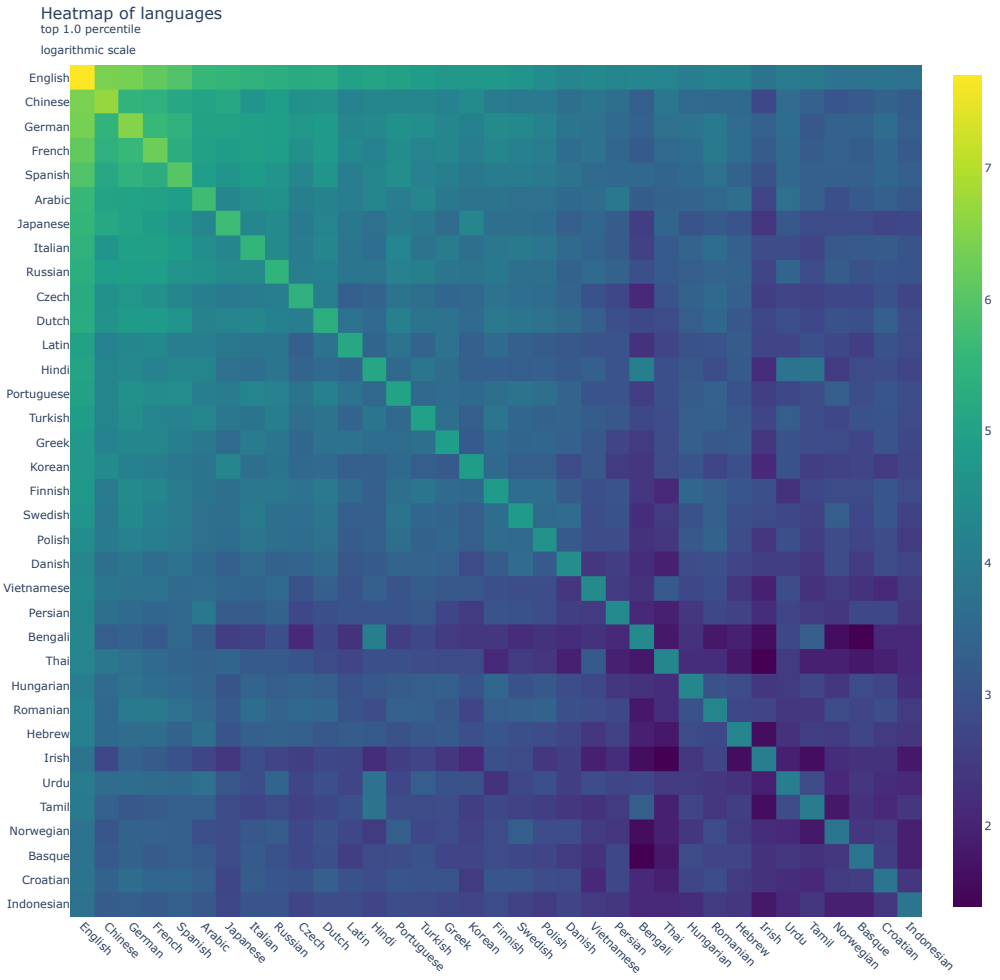

Figure 2: Heatmap of language co-occurrence in articles.

For clarity, the following is the ranking of top 20 most popular languages, by number of papers in which they have been considered:

1. English: 2215

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

2. Chinese: 809
3. German: 682
4. French: 533
5. Spanish: 416
6. Arabic: 306
7. Japanese: 299
8. Italian: 257
9. Russian: 239
10. Czech: 221
11. Dutch: 209
12. Latin: 171
13. Hindi: 166
14. Portuguese: 154
15. Turkish: 144
16. Greek: 133
17. Korean: 130
18. Finnish: 125
19. Swedish: 125
20. Polish: 98

As it is visible in Figure 2, the most popular language is English, but it may be caused by the bias of analyzing only English-language-written papers. Next, there is no particular positive, or negative, correlation between languages. However, there are slight negative correlations between languages Basque and Bengali, Irish and Thai, and Thai and Urdu, which means that these languages are very rarely researched together. There are two observations regarding these languages. (1) All of them are niche and do not have a big speaking population. (2) All pairs have very distant geographical origins, so there may be a low demand for their co-studying.

3.3. Findings related to RQ3: what are the popular fields, and topics, of research

Let us now discuss the finding related to the most popular fields and topics of reported research. In order to ascertain them, in addition to keyphrase search and NER, metadata mining and text summarization have been applied. Let us now introduce these methods in some detail.

3.3.1. Metadata mining

In addition to the information available within the text of a publication, further information can be found in its metadata. For instance, the date of publishing, overall categorization, hierarchical topic assignment and more, as discussed in the next paragraphs.

Therefore, metadata has been fetched both from the original source (arXiv API) and from the Semantic Scholar[17]. As a result, for each retrieved paper, the following information became available for further analysis:

- data: title, abstract and PDF,
- metadata: authors, arXiv category and publishing date,

[17] `https://www.semanticscholar.org`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

- citations/references,
- topics.

Note that the Semantic Scholar topics are different from the arXiv categories. The arXiv categories follow a set taxonomy[18], which is used by the person who uploads the text. On the other hand, the Semantic Scholar "uses machine language techniques to analyze publications and extract topic keywords that balance diversity, relevance, and coverage relative to our corpus."[19].

The metadata from both sources was complete for all articles (there were no missing fields for any of the papers). Obviously, one cannot guarantee that the information itself was correct. This had to be (and was) assumed, to use this data in further analysis.

3.3.2. Matching literature to research topics

In literature review, one may analyze all available information. However, it is much faster to initially check if a particular paper's topic is related to ones planned/ongoing research. Both Semantic Scholar and arXiv provide this information in the metadata. Semantic Scholar provides "topics", while arXiv provides "categories".

Figure 3 shows (1) what topics are the most popular (see the first column from the left), and (2) the correlation of topics. The measure used in the heatmap (correlation matrix) is the count of articles tagged with topics (logarithmic scale has been used).

Obviously, the most popular field of research is "Natural Language Processing". It is also worth mentioning that *Artificial intelligence*, *Machine Learning* and *Deep Learning* also score high in the article count. This is intuitive, as current applications of NLP are pursued using approaches from, broadly understood, artificial intelligence.

Moreover, the correlation, and high score, between "Deep Learning" and "Artificial Neural Networks" mirrors the influence of BERT and similar models. On the other hand, there are topics, which very rarely coincide. These are, for instance, *Parsing* and *Computer Vision*, *Convolutional Neural Networks* and *Machine Translation*, *Speech Recognition* and *Sentiment analysis*.

There is also one topic worth pointing out to: *Baseline (configuration management)*. According to the Semantic Scholar, it is defined as "an agreed description of the attributes of a product, at a point in time, which serves as a basis for defining change"[20]. This topic does not suit the NLP particularly, as it is too vague, and it could have been incorrectly assigned by the machine learning algorithm on the backend of Semantic Scholar.

Yet another interesting aspect is the evolution of topics in time, which gives a wider perspective of what topics are on the rise in, or fall from, popularity. Figures 4 show the most popular categories in time. The category cs.CL ("Computation and Language") is a dominating in all periods because it is the main subcategory of NLP. However, multiple interesting observation can be made. First, categories that are particularly popular nowadays are: cs.LG (Machine Learning), cs.AI (Artificial Intelligence), cs.CV (Computer Vision and Pattern Recognition). Second, there are categories, which experience a drop in interest. These are: stat.ML (Machine Learning) and cs.NE (Neural and Evolutionary Computing).

Moving to "categories" from arXiv, it is important to elaborate the difference between them and "topics". As mentioned, arXiv follows a taxonomy with two levels: primary category (always a single one) and secondary categories (may be many).

[18] `https://arxiv.org/category_taxonomy`
[19] `https://www.semanticscholar.org/faq\#extract-key-phrases`
[20] `https://www.semanticscholar.org/topic/Baseline-(configuration-management)/3403`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

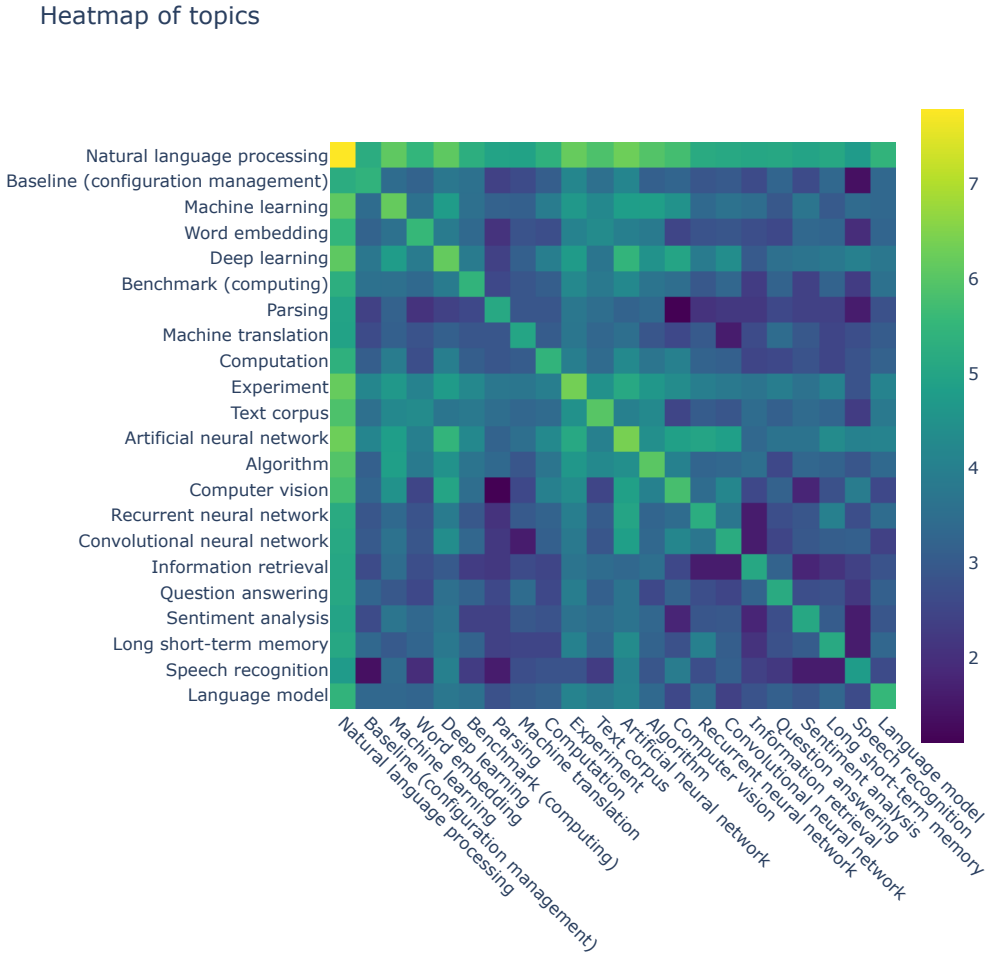

Figure 3: Correlation matrix between top 0.5 percentile of topics (logarithmic scale)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

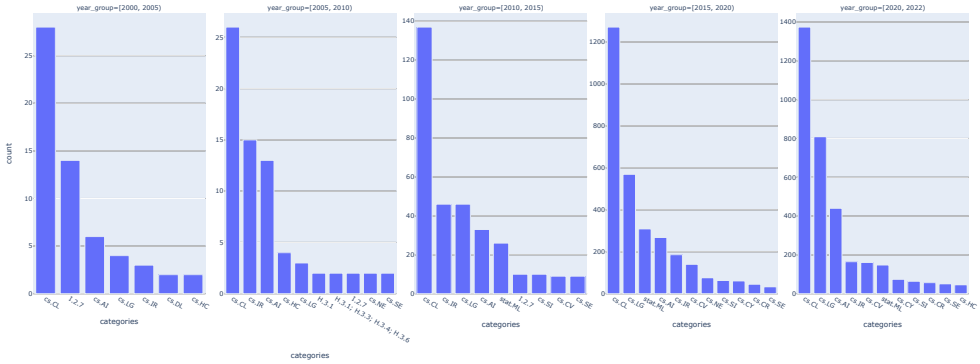

Figure 4: Most popular categories in time (top 96 percentile for each time period)

To best show this relation, as well as categories' popularity, a treemap chart has been created, which is most suitable for "nested" category visualization. It is shown in Figure 5.

Similarly to the Semantic Scholar "topics", the largest primary category is cs.CL (Computation and Language), which is a counterpart to the NLP topic from the arXiv nomenclature. Its top secondary categories are cs.LG/stat.ML (both categories of Machine Learning) and cs.AI (Artificial Intelligence). This is, again, consistent with previous findings and shows how these domains overlap each other. It is also worth noting the presence of cs.CV (Computer Vision and Pattern Recognition), which, although to a lesser degree, is also important in the NLP literature. Manual verification shows that, in this context, computer vision refers mostly to image description with text [95], visual question answering [96], using transformer neural networks for image recognition [97, 98], and other image pattern recognition, vaguely related to NLP.

Similarly, as for topics, a trend analysis has been performed for categories. It is presented in Figure 6. The most popular topic over time is *NLP*, followed by *Artificial neural network*, *Experiment*, *Deep learning*, and *Machine learning*. Here, no particular evolution is noticeable, except for rise in interests in the *Language model* topic.

3.3.3. Citations

Another interesting metainformation, is the citation count [84, 85]. Hence, this statistic was used to determine key works, which were then used to establish key research topic in NLP (addressing also RQ1-3).

It is well known that, in most cases, the distribution of node degree in a citation network is exponential [94]. Specifically, there are many works with 0-1 citations, and very few with more than 10 citations. In this context, the citations network of top 10% of most highly cited papers is depicted in Figure 7. The most cited papers are 1810.04805v2 [31] (5760 citations), 1603.04467v2 [86] (2653 citations) and 1606.05250v3 [87] (1789 citations). The first one is the introduction of the BERT model. Here, it is easy to notice that this papers absolutely dominates the network in terms of the degree. It is the networks focal point. This means that the whole domain not only revolves around one particular topic, but also around a single paper.

The second paper concerns TensorFlow, the state-of-the-art library for neural networks construction and management. The third introduces "Squad" – a text dataset with over 100,000 questions, used for machine learning. It is important to note that these are the top 3 papers when

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

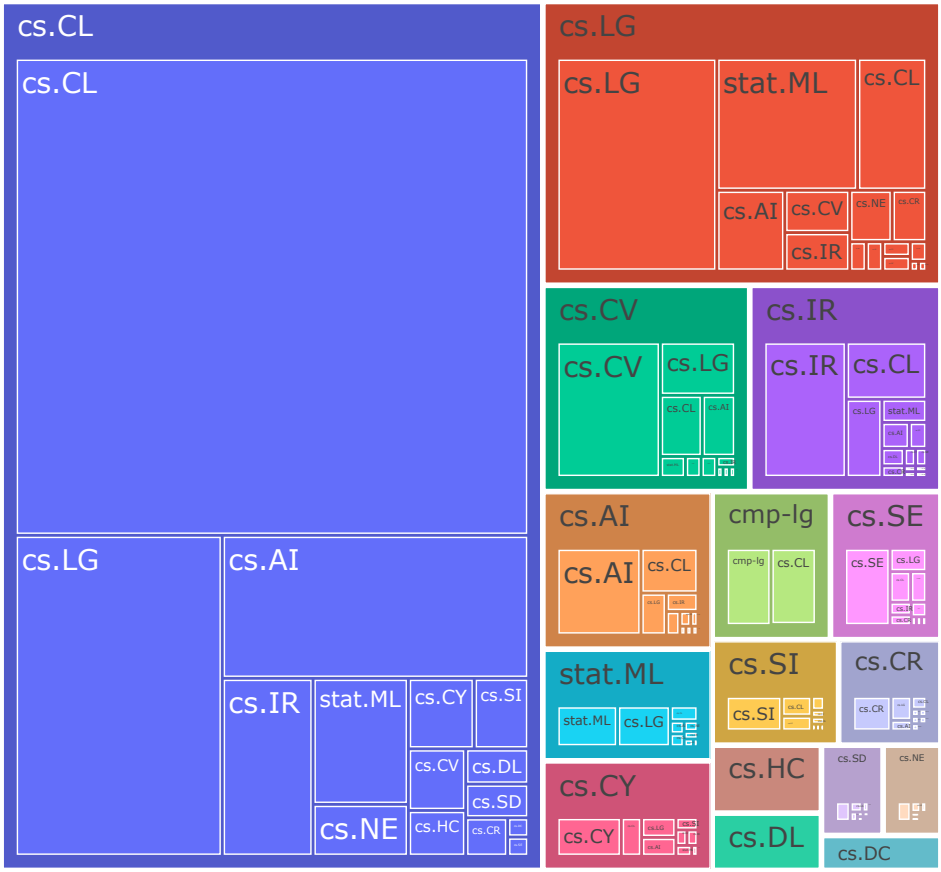

Figure 5: Simplified treemap visualizing arXiv primary categories aggregating secondary categories. Outer rectangles are primary categories, inner rectangles are other assigned categories. Other categories include primary category to additionally show the primary categories size. Top 20.0 of primary categories and categories. Colors are purely aesthetic.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

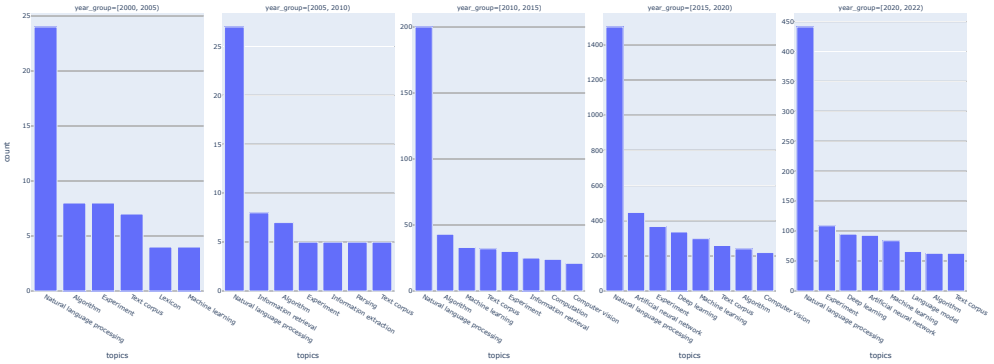

Figure 6: Most popular topics in time (top 99,8 percentile for each time period)

considering not only works published after 2015, but also when the "all time most cited works" are searched for.

How can two papers cite each other. An interesting observation has been made, during the citation analysis. Typically, relation, where one paper quotes another paper, should be one-way. In other words when paper A *cites* paper B, that means that paper B is a reference for paper A. So the set of citations and reference should be disjoint. This is true for over 95% of works. However, 363 of papers have an intersection between citations and references, with the biggest having even 10 common positions. Further, manual, analysis determined that this "anomaly" happens due to the existence of preprints, and all other cases where a paper appeared publicly (e.g. being a Technical Report) and then was revised and cited a different paper. This may happen, for instance, when a paper is criticised and it is reprinted (an updated version is created) to address the critique.

3.4. RQ3 related findings based on application of keyphrase and entity networks

As discussed, NER has been used to determine NLP datasets and languages analyzed in papers. It can also be used when looking for techniques used in research. However, to better visualize the topic of interest, it can be combined with network analysis. Specifically, work reported in the literature involves many-to-many relations, which provide information of what techniques, methods, problems, languages etc., are used alone, in tandem or, perhaps, in groups. To properly explore the area, four dimensional networks (see Figures 8 and 9) have been constructed, with: nodes (entities), node size (scaled by an attribute), edges (relations), edge width (scaled by an attribute). Moreover, since all networks are exponential and have very high edge density, only the top percentile of entities has been graphically represented (to allow readability). Networks have been built using networkx [83] and igraph [82] Python libraries.

As shown in the Figure 8 the majority of entities are related to models such as BERT, and neural network architectures (e.g. RNN, CNN). However, the findings show not only NLP-related topics, but all entities. Here, an important warning, regarding used NER models, should be stated. In most cases, when NER is applied directly, and without additional techniques, the entities are not disambiguated, or properly unified. For instance, surnames, like, Kim, Wang, Liang, Liu, Chen, etc. are not properly recognized as names of different persons and "bagged

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

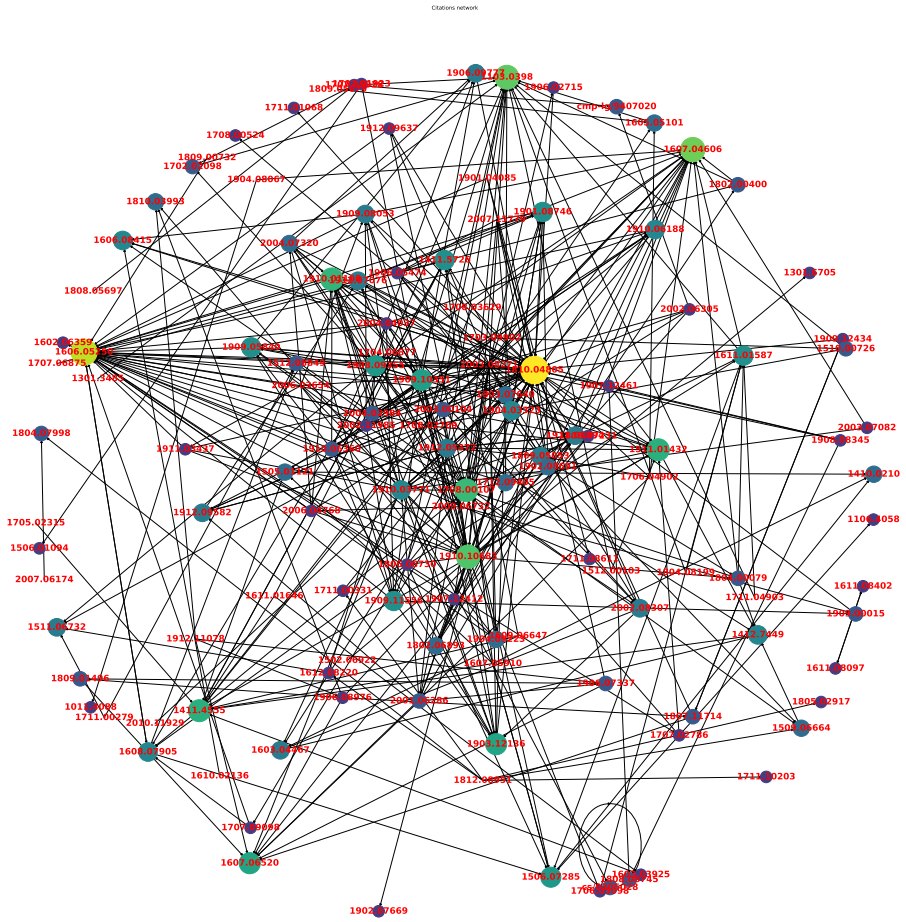

Figure 7: Citation network of all articles (arrows point towards cited paper); top 5 percentile; A → B, means A cites B (B is a reference of A); Color scale indicates how many papers cite a given paper (yellow – higher, dark blue – lower)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

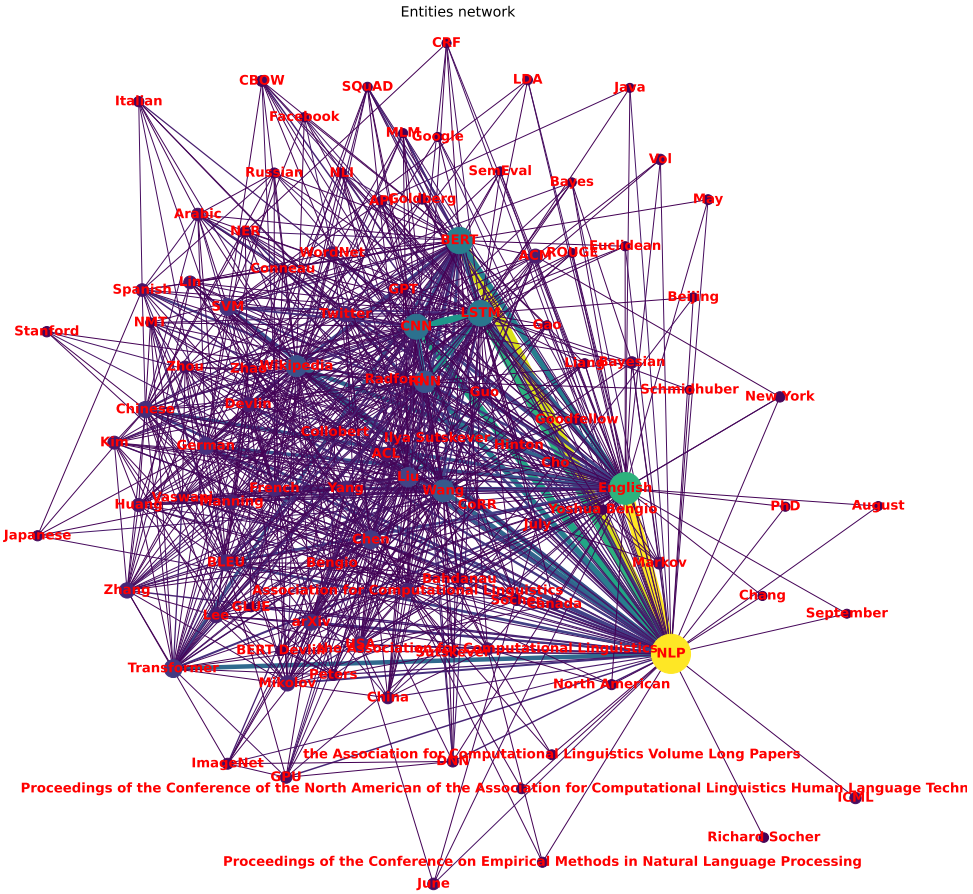

Figure 8: Entity network; entities detected using spaCy (en_core_web_lg 3.1.0); edges width – number of papers with the entity; node size and color – sum of weight of edges; top 0.4 percentile of node weight; top 20.0 percentile of edge weight

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

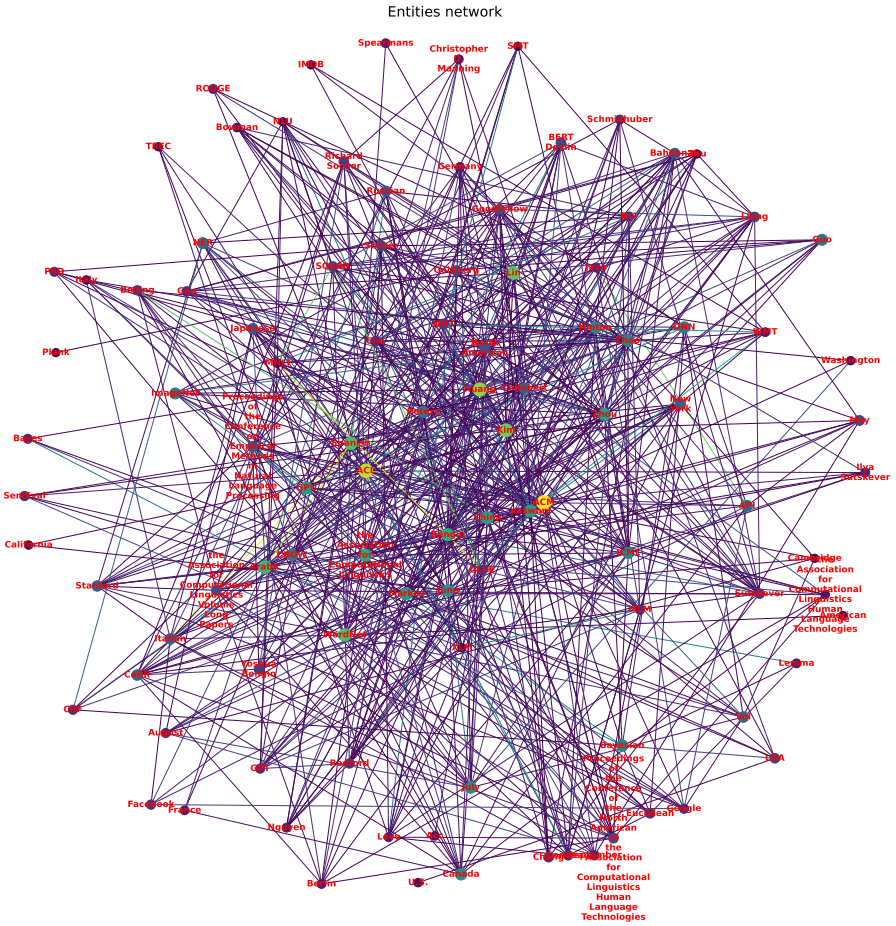

Figure 9: Entity network; entities detected using spaCy (en_core_web_lg 3.1.0); edges width – number of papers with the entity; node size and color – sum of weight of edges; node weight between 99.5 and 99.9 percentile; top 20.0 percentile of edge weight

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

together". Therefore, further interpretation of results of NER may require manual checking of results.

Moreover, corroborating earlier noted result, is that Wikipedia and Twitter, being the most popular data sources for NLP, can be observed.

Finally, among important entities, Association for Computational Linguistics (also shown as "the Association for Computational Linguistics" and "ACL"[21]) has been found. This society organizes conferences, events and also runs a journal about natural language processing.

Figure 9 shows very popular name entities, but skips the most often found ones. This has been done to allow other frequent terms to become visible. Specifically, the networks were trimmed by node weight, i.e. number of papers including the named entity. The Figure 9 contains terms between the 99.5 and 99.9 percentiles by node weight. In addition to some previously made observations, new entities appeared, which show what is also of considerable interest in NLP literature. These are:

- GPU (Graphic Processing Unit), which are often used to accelerate neural network training (and use) [103]
- WordNet – semantic network "connecting words" with regard to their meaning [104][22] and ImageNet – a image database using WordNet hierarchy to propose a network of images [108][23]
- SemEval – popular contents in NLP, occurring annually and challenging scientist with different NLP tasks [24]
- and other particular methods like (citation contain example papers): Bayesian methods [105], CBOW (Continuous Bag of Words) [106], Markov processes [107]

As described in Section 3.1.2, the keyphrase search was used to extract these terms and findings, which might have been skipped in the NER results. For example, the word "accuracy" is a widely used metric in NLP and many other domains. However, it is not a named entity, because it is also an "ordinary" word in English and is not detected as such by the NER models. Applied analysis produced a network of keyphrase co-occurrence. Hence, network visualization was, again, applied (Figure 10). This allowed formulation of hypotheses, which underwent further (positive) manual verification, specifically:

- BERT models are most commonly used in their pretrained "version" / "state". BERT is already a pretrained model, but it is possible to continue its training (to get a better representation of particular language, topic or domain). The second approach is using BERT, or its pretrained variant, to train it on a target task, called downstream task (these techniques is also called "fine-tuning").
- Transformers are connected strongly with attention. This is because transformer (a neural network architecture) is characterized by the presence of attention mechanism in it. This is the distinguishing factor of this architecture [161].

[21] https://www.aclweb.org
[22] https://wordnet.princeton.edu
[23] https://image-net.org
[24] https://semeval.github.io

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

- "Music" is connected with "lyrics". This shows that the intersection between NLP research and music domain is via lyrics analysis. The lack of correlation between music and other terms shows that audio analysis, sentiment analysis, etc. are not that popular in this context.
- "Precision" is connected with "recall" These two extremely popular evaluation metrics for classification are often used together. Their main point is to handle imbalanced datasets, where the performance is not evaluated correctly by the "accuracy" [140] measure.
- "Synset" is connected with "WordNet". As shown, WordNet is most commonly used with Synset (a user programmer-friendly interface available in the NLTK framework[25]).
- Quantum mechanics begins to emerge in NLP. The oldest works in the field of quantum computing (in the set under study) date back to 2013 [134], but most (>90%) of the recent works dates to 2019-2021. These provide answers to the to problems such as: applying NLP algorithm on "nearly quantum" computers [138], sentence meaning inference with quantum circuit model(s), encoding-decoding [135], quantum machine learning [136] or, even, ready-to-use Python libraries for the quantum NLP [137] are investigated. There are still very few works joining the worlds of NLP and quantum computing, but their number is significantly growing since 2019.
- Graphs are very common in research related to semantic analysis. One of the the domains that NLP overlaps/includes is semantics. The entities network illustrates how important the concept of a graph is in semantics research (e.g. knowledge graphs). Some works touch these topics in tandem with text embedding [141], text summarization [142], knowledge extraction/inference/infusion [142] or question answering [143].

3.4.1. Text summarization

Another approach to extract key information (including the field of research) is to reduce the original text to a brief and simple "conclusion". This can be done with *extractive* and *abstractive* summarization methods. Both aim at allowing the user to comprehend the main message of the text. Moreover, depending on what sentences are chosen in the extractive summarization methods, one may find which abstracts (and papers) are most "summaritive".

Extractive summarization. First, the extractive methods have been used to summarize the text of all abstracts. Specifically, the following methods have been applied.

- Luhn methods [60] (max 5 sentences) shown in Listing 1
- Latent Semantic Analysis [61] (max 5 sentence) shown in Listing 2
- LexRank [63] (max 5 sentence) shown in Listing 3
- TextRank [62] (max 5 sentence) shown in Listing 4

Here, note that, due to formatting errors in the original texts, the library pysummarization[26] had trouble with "sentences with periods" (e.g. "3.5% by the two models, respectively." is only a part of a full sentence, but it contains a period character).

[25] `https://www.nltk.org/howto/wordnet.html`
[26] `https://pypi.org/project/pysummarization`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

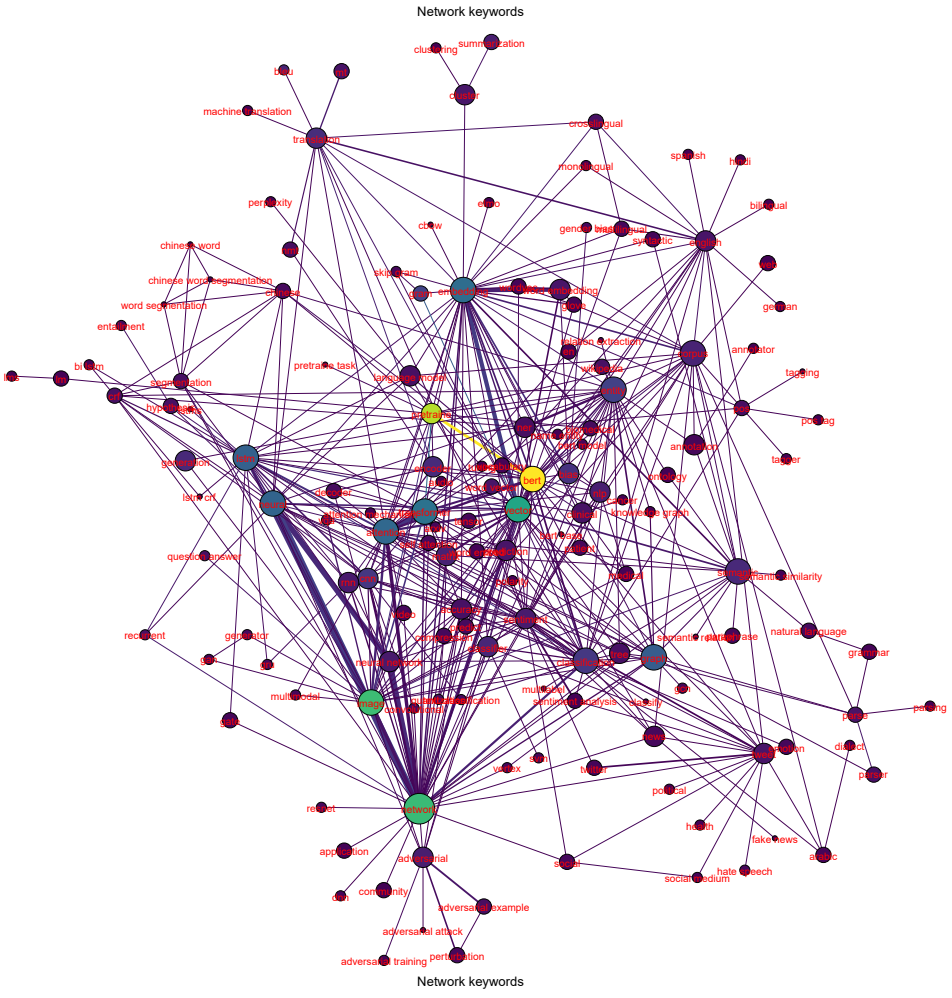

Figure 10: Keyphrase co-occurrence network Node size – article count where keyword appears Node color – citation sum where keyword appears Edge width & color – number of articles in which two terms appeared

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Abstractive summarization. Previous research found that abstractive summarization methods can "understand the sense" of the text, and build its summary [58]. It was also found that their overall performance is better than that of extractive methods [57]. However, most state-of-the-art solutions have limitations related to the maximum number of tokens, i.e. BERT-like models (e.g. distilbart-cnn-12-6 model [65], bart-large-cnn [65], bert-extractive-summarizer [59]) support maximum of 512 tokens, while the largest Pegasus model supports 1024 [68].

Nevertheless, very recent work proposes a transformer model for long text summarization, a "Longformer" [67], which is designed to summarize texts of 4000 tokens and more. However, this capability comes with a high RAM memory requirement. So, in order to test abstractive methods, Longformer was applied only to titles of most influential texts (top 5% of citation count).

The final note about text summarization is that, most recent research proposed innovative ways to overcome the length issue (see, [66]). There is thus a possibility to apply text summarization, for instance, to abstracts combined with introduction and conclusions of research papers. Testing this possibility may be a good starting point for research, but is out of scope of this contribution.

3.4.2. Summarization findings

Listings 1, 2, 3, 4, show summaries of all abstracts and Listing 5 shows summary of all titles (as described in Section 3.4.1).

The common part for all summaries addresses (in a hierarchical order, starting from most popular features):

- natural language processing and artificial intelligence,
- translation and image processing,
- neural networks,
- deep neural network architectures, e.g. CNN, RNN, encoder-decoder, transformers, and
- deep neural network models, e.g. BERT, ELMO.

Moreover, the main "ideas", which appear in the summaries are: effectiveness, "state-of-the-art" solutions, and solutions "better than others". This shows the "competitive" and "progress-focused" nature of the domain. Authors find it necessary to highlight how "good" or "better than" their solution is. It may also mean that there is not much space for "exploratory" and "non-results-oriented" research (at least this is the message permeates the top cited articles). Similarly, research indicating which approaches do not work in a given domain is not appreciated.

Listing 1: Text summarization with method Luhn

```
Summary with LSA (512.9 sec)
Natural language processing, as a data analytics related
    technology, is used widely in many research areas such as
    artificial intelligence, human language processing, and
    translation. [paper id: 1608.04434v1]
At present, due to explosive growth of data, there are many
    challenges for natural language processing. [paper id:
    1608.04434v1]
```

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213


```
Hadoop is one of the platforms that can process the large
    amount of data required for natural language processing. [
    paper id: 1608.04434v1]
KOSHIK is one of the natural language processing architectures,
     and utilizes Hadoop and contains language processing
    components such as Stanford CoreNLP and OpenNLP. [paper id:
    1608.04434v1]
This study describes how to build a KOSHIK platform with the
    relevant tools, and provides the steps to analyze wiki data.
     [paper id: 1608.04434v1]
```

Listing 2: Text summarization with method Latent Semantic

```
Summary with sumy-LSA (512.9 sec)
Natural language processing, as a data analytics related
    technology, is used widely in many research areas such as
    artificial intelligence, human language processing, and
    translation. [paper id: 1608.04434v1]
At present, due to explosive growth of data, there are many
    challenges for natural language processing. [paper id:
    1608.04434v1]
Hadoop is one of the platforms that can process the large
    amount of data required for natural language processing. [
    paper id: 1608.04434v1]
KOSHIK is one of the natural language processing architectures,
     and utilizes Hadoop and contains language processing
    components such as Stanford CoreNLP and OpenNLP. [paper id:
    1608.04434v1]
This study describes how to build a KOSHIK platform with the
    relevant tools, and provides the steps to analyze wiki data.
     [paper id: 1608.04434v1]
```

Listing 3: Text summarization with method LexRank

```
Summary with LexRank (11323.26 sec)
Many natural language processing applications use language
    models to generate text. [paper id: 1511.06732v7]
However, there is no known natural language processing (NLP)
    work on this language. [paper id: 1912.03444v1]
However, few have been presented in the natural language
    process domain. [paper id: 2107.07114v1]
Here, we show their effectiveness in natural language
    processing. [paper id: 2109.04712v1]
The other two methods however, are not as useful. [paper id:
    2109.01411v1]
```

Listing 4: Text summarization with method TextRank

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Summary with sumy-TextRank (497.67 sec)

Recently, neural models pretrained on a language modeling task, such as ELMo (Peters et al., 2017), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2018), have achieved impressive results on various natural language processing tasks such as question-answering and natural language inference. [paper id: 1901.04085v5]

In chapter 1, we give a brief introduction of the history and the current landscape of collaborative filtering and ranking; chapter 2 we first talk about pointwise collaborative filtering problem with graph information, and how our proposed new method can encode very deep graph information which helps four existing graph collaborative filtering algorithms; chapter 3 is on the pairwise approach for collaborative ranking and how we speed up the algorithm to near-linear time complexity; chapter 4 is on the new listwise approach for collaborative ranking and how the listwise approach is a better choice of loss for both explicit and implicit feedback over pointwise and pairwise loss; chapter 5 is about the new regularization technique Stochastic Shared Embeddings (SSE) we proposed for embedding layers and how it is both theoretically sound and empirically effectively for 6 different tasks across recommendation and natural language processing; chapter 6 is how we introduce personalization for the state-of-the-art sequential recommendation model with the help of SSE, which plays an important role in preventing our personalized model from overfitting to the training data; chapter 7, we summarize what we have achieved so far and predict what the future directions can be; chapter 8 is the appendix to all the chapters. [paper id: 2002.12312v1]

We explore how well the model performs on several languages across several tasks: a diagnostic classification probing the embeddings for a particular syntactic property, a cloze task testing the language modelling ability to fill in gaps in a sentence, and a natural language generation task testing for the ability to produce coherent text fitting a given context. [paper id: 1910.03806v1]

Neural Architecture Search (NAS) methods, which automatically learn entire neural model or individual neural cell architectures, have recently achieved competitive or state-of-the-art (SOTA) performance on variety of natural language processing and computer vision tasks, including language modeling, natural language inference, and image classification. [paper id: 2010.04249v1]

Transfer learning in natural language processing (NLP), as realized using models like BERT (Bi-directional Encoder

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Listing 5: Title summarization (top 5 percentile papers) with Longformer

```
'The Natural Language Processing (NLT) is a new tool that can
    teach people about the world. The tool is based on the data
    collected by CNN and RNN. A survey of the Usages of Deep
    Learning was carried out by the 2015 MSCOCO Image Search. It
     was created by a survey of people in the UK and the US. An
    image is worth 16x16 words, and a survey reveals how many
    people are interested in the language.'
```

```
    Representation from Transformer), has significantly improved
     language representation with models that can tackle
    challenging language problems. [paper id: 2104.08335v1]
```

3.5. RQ1, RQ2, RQ3: Relations between NLP datasets, languages, and topics of research

Additionally, to separate results for RQ1, RQ2 and RQ3, there are situations when important information is the coincidence of these three aspects: NLP datasets, languages, and research topics. The triplet dataset-language-problem is usually fixed on two positions. For example, a research may be focused on machine translation (problem) into English (language), but with missing a corpus (dataset); or a group of Chinese researchers (language) has access to a rich Twitter API (dataset), but is considering what type of analysis (problem) is most prominent. This sparks a question what datasets are used, with which languages, and for what problems. Presented results of correlations between these 3 aspects are divided into two groups, for 2 most popular language: English and Chinese. They are shown in Figure 11. The remaining results for the selected languages, from the most popular ones, can be found in Figure 12 and 13.

For English and Chinese languages (being the subject of NLP research) the distribution of problems is very similar. The top problems are: machine translation, question answering, sentiment analysis and summarization. The most popular dataset used for all of these problems is Wikipedia. Additionally, for sentiment analysis, there is a significant number of contributions that use also Twitter. All of these observations are consistent with previous results (reported in Sections 3.1 3.6 3.2).

Before going into languages other than English and Chinese, it is crucial to recall that this analysis focused only on articles written in English. Hence, reported results may be biased in the case of research devoted to other language(s). Nevertheless, there exists a large body of work about NLP applied to non-English languages, which is written in English. For instance, among all analyzed papers for this contribution, 41% were devoted to NLP in the context of neither English (non-english papers are 46% of the dataset) nor Chinese (non-chinese papers are 80% of the dataset).

The most important observation is that the distribution of problems for languages other than English and Chinese is, overall, similar (Machine Translation, Question-Answering, sentiment and summarization are the most popular ones). However, there are also some distinguishable differences:

- For German and French, summarization, language modelling and natural language inference, and named entity recognition are the key research areas.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Figure 11: Datasets and NLP problems for languages English and Chinese.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

- In Arabic and Italian, Japanese, Polish, Estonian, Swedish and Finish, there is a visible trend of interest in named entity recognition.
- Dependency parsing is more pronounced in research on languages such as German, French, Czech, Japanese, Spanish, Slovene, Swahili and Russian.
- In Basque, Ukrainian, Bulgarian the domain does not have particular homogeneous sub-domain distribution. The problems of interests are: co-reference resolution, dependency parsing, dialogue-focused research, language modeling, machine translation, multitask learning, named entity recognition, natural language inference, part-of-speech tagging, question answering.
- In Bengali, a special area of interest is part-of-speech tagging.
- Research focused on Catalan have a particular interests in dialogue-related texts.
- Research regarding Indonesian have a very high percent of sentiment analysis research. Even higher than most popular topic of machine translation.
- Studies on Norwegian language are strongly focused on sentiment analysis, which peeks over the most common domain of most of the languages – machine translation.
- Research focusing on Russian puts a special effort in analyzing dialogues and dependency parsing.

There are only minimal difference between datasets used for English and Chinese, and other languages. The key ones are:

- Facebook is present as one of the main sources in many languages, being particularly popular data source for: Bengali, and Spanish
- Twitter is a key data source in research on languages: Arabic, Dutch, French, German, Hindi, Italian, Korean, Spanish, Tamil
- WordNet is very often used in research involving: Moldovan and Romanian
- Tibetan language research nearly never uses Twitter as the dataset.

3.6. Findings concerning RQ4: most popular specific tasks and problems

At the heart of the research is yet another key aspect – the specific problem that is being tackled, or the task, which is being solved. This may seem similar to the domain, or to the general direction of the research. However, some general problems contain specific problems (e.g. machine translation and English-Chinese machine translation, or named entity recognition and named entity linking). On the other hand, some specific problems have more complicated relation, e.g. machine translation, which in NLP can be solved using neural networks, but neural networks are also an independent domain on their own, which is also a superdomain (or a sub-domain) of, for instance, image recognition. These complicated relations point to the need for a standardized NLP taxonomy. This, however, is also out of scope of this contribution.

Let us come back to the methods of analyzing specific results. To extract most popular specific tasks and particular problems, methods described above, such as NER, keyphrase search,

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Figure 12: Datasets and NLP problems for chosen languages.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

Figure 13: Datasets and NLP problems for chosen languages.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

metadata mining, text summarization, and network visualization were used. Before presenting specific results, an important aspect of keyphrase search needs to be mentioned. An unsupervised search for particular specific topics of research cannot be reasonably performed. All approaches of unsupervised keyphrase search that have been tried (in an exploratory fashion) produced thousands of potential results. Therefore, supervised keyphrase search has been applied. Therefore, the NLP problems were determined based on an exhaustive (multilingual) list, aggregating most popular NLP tasks[27].

The list has been extracted from the website and pruned of any additional markdown[28], to obtain a clean text format. Next, all keywords and keyphrases from the text of each paper has been compared with the NLP tasks list. Finally, each paper has been assigned a list of problems found in its text. Figure 14 shows the popularity (by count) of problems addressed in NLP literature.

Again, there is a dominating problem – machine translation. This is very intuitive, if one takes into account the recent studies [119, 121, 122, 123, 124] showing that lack of high fidelity machine translation remains the key barrier for world-wide communication. This problem seems very persistent, because it was indicated also in older research (e.g. in text from 1968 [120]). Here, it is important to recall that this contribution is likely to be biased towards translation involving English language, because it only analyzed English-written literature.

The remaining top 3 most popular problems are question answering [126] and sentiment analysis [129]. In both these domains, there are already state-of-the-art models ready to be used[29]. What is interesting, for both question answering and sentiment analysis, most of the models are based either on BERT or its variation, DistilBERT [125].

3.7. RQ5: seeking outliers in the NLP domain

Some scientific research areas are homogeneous, and all publication revolve around similar topic (group of topics). On the other hand, some can be very diverse, with individual papers touching very different subfields. Finally, there are also domains where, from a more or less homogeneous set, a separate, distinguishable, subset can be pointed to. To verify the structure of the field of NLP, two methods have been used. One is, previously introduced, metadata mining. The second one was text embedding and cauterization. Let us briefly introduce the second one.

3.7.1. Text embeddings

One of ubiquitous methods in text processing are word, sentence and document embeddings. Text embeddings, which ”convert texts to numbers”, have been used to determine key differences/similarities between analyzed texts.

Embeddings can be divided into: contextualized and context-less [33]. Scientific papers often use words, which strongly depend on the context The prime example is the word “BERT” [31], which on the one hand is a character from a TV show, but in the NLP world it is a name of one of the state-of-the-art embedding models. In this context, envision application of BERT, the NLP method, to analysis of dialogues in children TV, where one of the dialogues would include BERT, the character. Similar situation concerns words like network (either neural network, graph network, social network, or computer network), “spark” [32] (either a small fiery particle,

[27] https://github.com/sebastianruder/NLP-progress
[28] https://www.markdownguide.org
[29] https://huggingface.co/models?language=en&pipeline_tag=question-answering

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

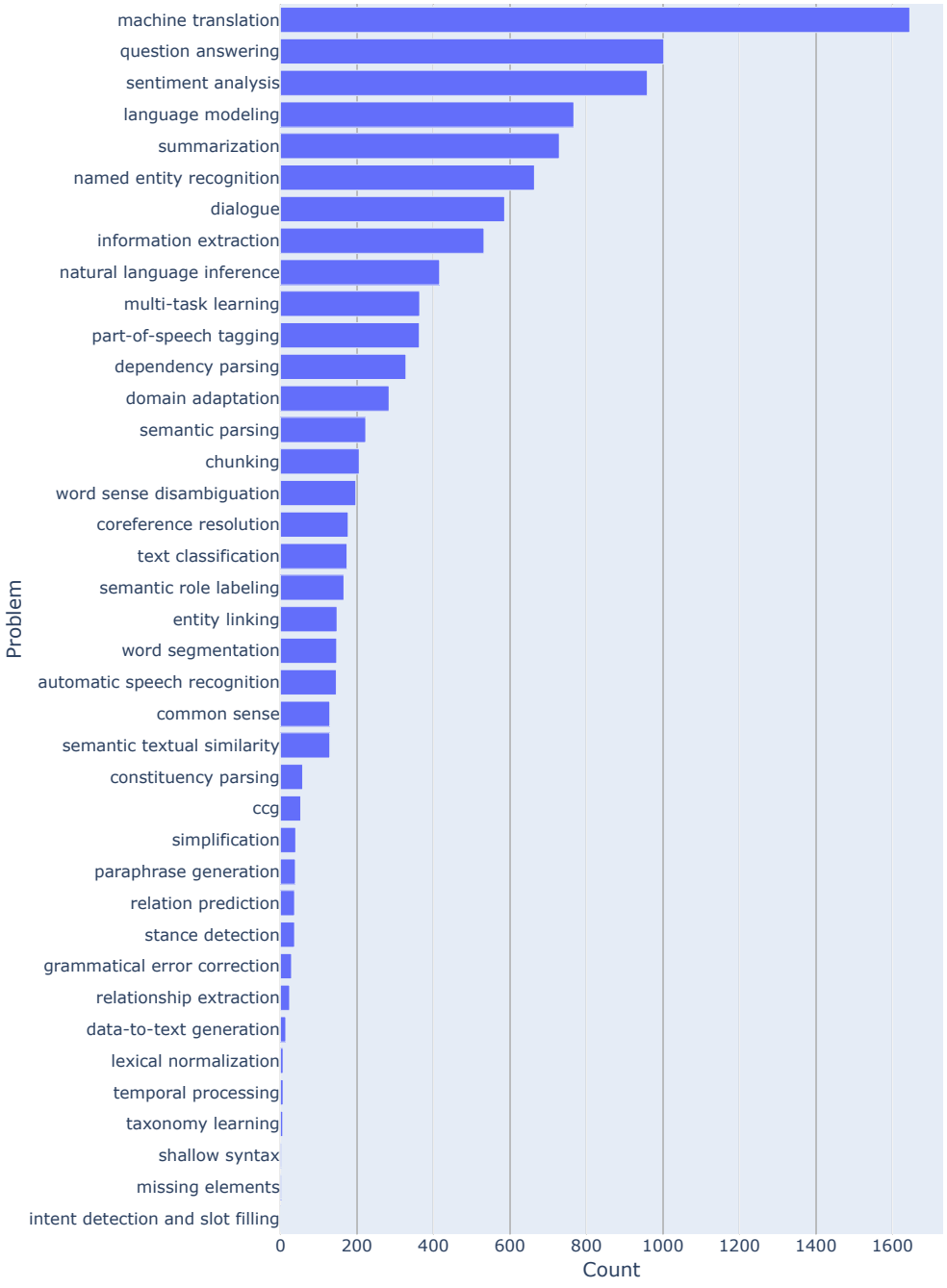

Figure 14: Histogram of problems tackled in NLP literature.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

or the name of a popular Big Data library), lemma (either a proven proposition in logic, or a morphological form of a word), etc. Hence, in this study, using contextualized text embeddings is more appropriate. This being the case, very popular static text embeddings like Glove [35] and Word2Vec [36, 37] have not been used.

There are many libraries and models available for contextualized text embedding, e.g.: transformers [101], flair [34], gensim [159] and models: BERT [31] (and its variations like Roberta [39], DistilBERT [125]), GPT-2 [38], T5 [40], ELMo [99] and others. However, most of them require specific and high-end hardware to operate reasonably fast (i.e. GPU acceleration [29]). Here, the decision was to proceed with FastText [30]. FastText is designed to produce time efficient results, which can be recreated on standard hardware. Moreover, it is designed for "text representations and text classifiers"[30], which is exactly what is needed in this work.

3.7.2. Embedding and clustering

It is important to highlight that since FastText, like most embeddings, has been trained on a pretty noisy data [30], the input text of articles was preprocessed only with Stage 1 cleaning (see Section 2.2). Next, a grid search [53] was performed, to tune hyperparameters. While, as noted earlier, hyperparameter tuning has not been applied, use of grid search, reported here, illustrates that there exist ready-to-use libraries that can be applied when hyperparameter tuning is required. Overall, the best embeddings were produced by a model with the following hyperparameters[31]:

- dimension: 20
- minimum subword size: 3
- maximum subword size: 6
- number of epochs: 5
- learning rate: 0.00005

Finally, the FastText model was further trained in an unsupervised mode (which is standard in majority of cases for general language modelling), on texts of papers, to better fit the representation.

After embeddings have been calculated, their vector representations have been clustered. Since there was no response variable, an unsupervised classifier was applied. Again (as in Section 3.7.1), the main goal was simplicity and time efficiency.

Out of all tested algorithms (K-means [45], OPTICS [46, 47], DBSCAN [48, 49], HDBSCAN [51] and Birch [52]), the best time efficiency, combined with relative simplicity of use, was achieved with K-means (see, also [42, 41]). Moreover, in found research, K-means clustering showed best results, when applied to FastText embeddings (see, [170]).

The evaluation of clustering has been performed using three clustering metrics: Silhouette score [54], Davies-Bouldin score [55], Caliński-Harabasz Score [56]. These metrics were chosen because they allow evaluation of unsupervised clustering. To visualize the results on a 2D plane, the multidimensional FastText vectors were converted with t-distributed stochastic neighbor embedding (T-SNE) method [44, 43]. T-SNE has been suggested by text embedding visualizations reported in earlier work [173, 172].

[30] `https://fasttext.cc`
[31] `https://fasttext.cc/docs/en/options.html`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

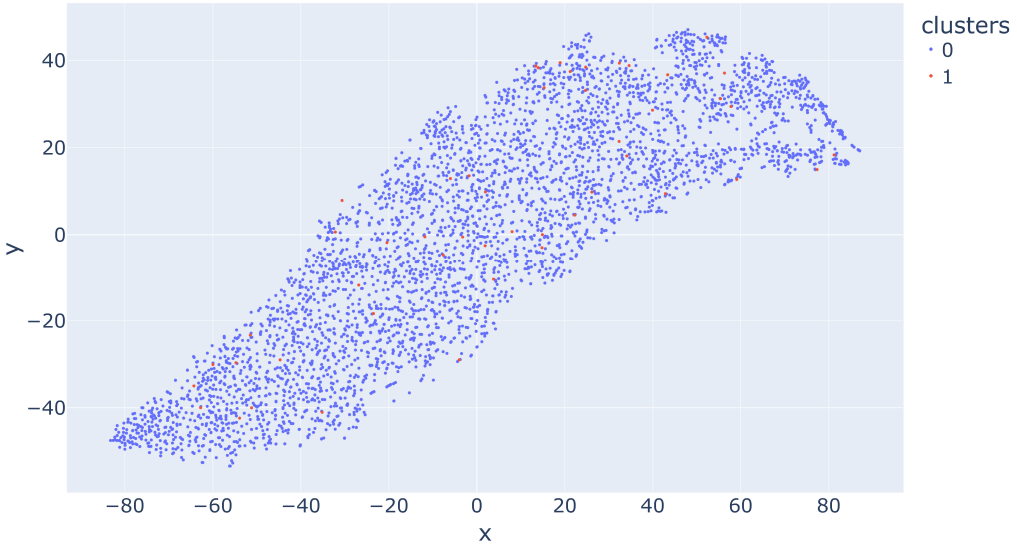

Figure 15: "The blade of NLP". A visualization of all paper text embeddings grouped in clusters (dimensionality reduced with T-SNE).

3.7.3. RQ5: outliers found in the NLP research

Visualizations of embeddings are shown in Figure 15.

Note that Figure 15 is mainly aesthetic, as actual relations are rarely visible, when dimension reduction is applied. The number of clusters has been evaluated according to 3 clustering metrics (Silhouette score [54], Davies-Bouldin score [55], Caliński-Harabasz Score [56]) and the best clustering score has been achieved for 2 clusters. Hence, further analysis considers separation of the embeddings into 2 clusters. To further explore why these particular embeddings appear in the same group, various tests were performed. First, wordclouds of texts (titles and paper texts) in the clusters have been built. The texts for wordclouds were processed with Stage 2 cleaning. Title wordclouds are shown in Figure 16, while text wordclouds are shown in Figure 17.

Further, citation count comparison (Figures 18 and 19) and authors were checked for text in both clusters.

Last, the differences in topics from Semantic Scholar (Figures 20 and 21) and categories from arXiv (Figures 22 and 23) have been checked.

Based on the content of Figures 16, 17, 18, 19, 20, 21, 22, 23 and the author per cluster distribution analysis the following conclusions have been drawn:

- There is one specific outlier, this is the cluster of work related to texts embeddings.
- Content of texts shows strong topical shift towards deep neural networks.
- Categories and topics of clusters are not particularly far away from each other, because their distribution is similar. There is a higher representation of computer vision and information retrieval area in the smaller cluster (cluster 0).
- There are no distinguishable authors who are responsible for texts in both clusters.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

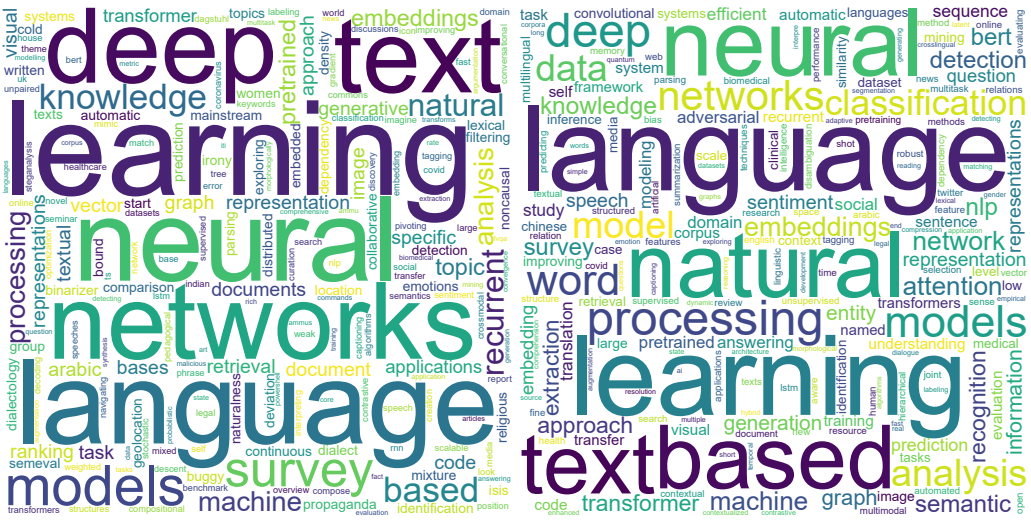

Figure 16: Wordcloud of titles of papers in cluster 0 (left) and 1 (right)

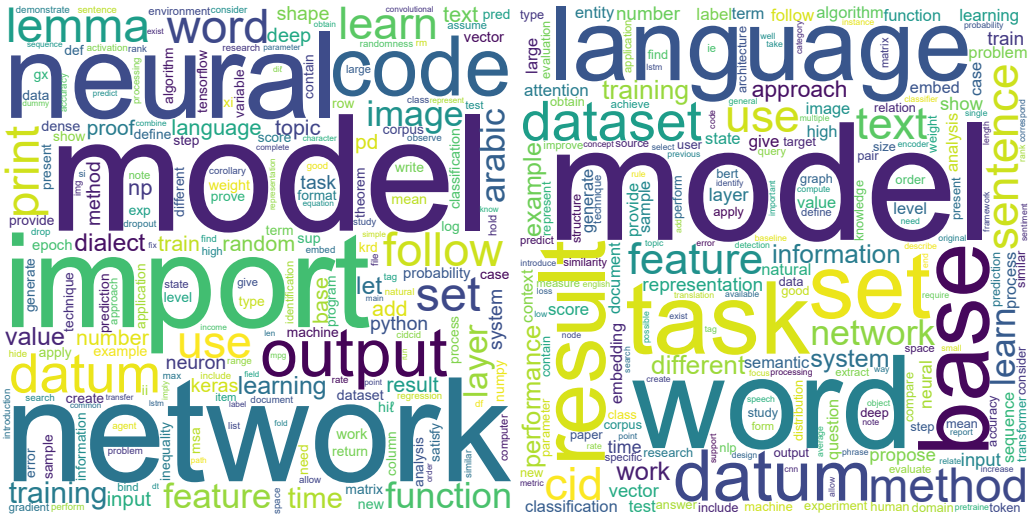

Figure 17: Wordcloud of texts of papers in cluster 0 (left) and 1 (right)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

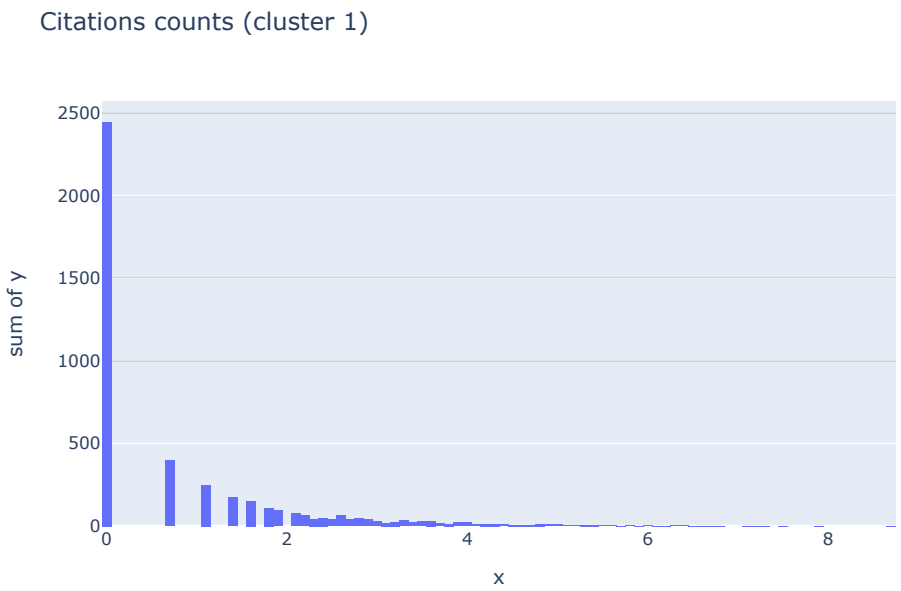

Figure 18: Histogram of citation counts in cluster 1 (bigger cluster) - logarithmic scale

- The distribution of citation counts is similar in both clusters.

Furthermore, manual verification showed that deep neural networks is actually the biggest subdomain of NLP, and it touches upon issues, which do not appear in other works. These issues are strictly related to neural networks (e.g. attention mechanism, network architectures, transfer learning, etc.) They are universal, and their applications play an important role in NLP, but also in other domains (image processing[109], signal processing [110], anomaly detection [112], clinical medicine [113] and many others [111]).

3.7.4. "Most original papers"

In addition to unsupervised clustering, an additional approach to outlier detection has been applied. Specifically, metadata representing citations/reference information was further analyzed. On the one hand, of the "citation spectrum" are the most influential works (as shown in Section 3.3.3). On the other side, there are papers that either are new and have not been cited yet, or those that do not have high influence. However, the true "original" works are papers which have many citations (they are in top 2 percentile), but very few references (bottom 2 percentile). Based on performed analysis, it was found that such papers are:

- "Natural Language Processing (almost) from Scratch" [88] – a neural network approach to learning internal representations of text, based on unlabeled training data. A similar idea was used in future publications, especially, the most cited paper about BERT model [31].

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

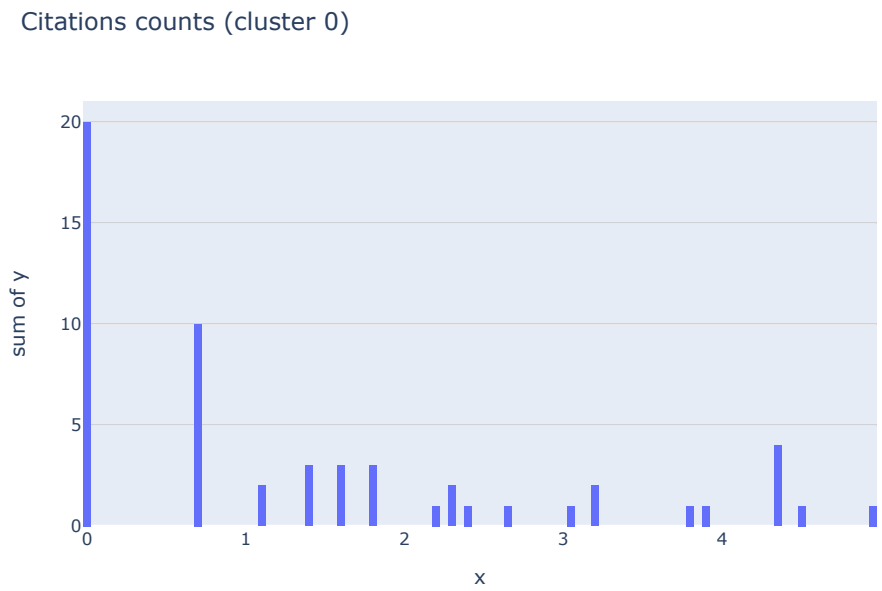

Figure 19: Histogram of citation counts in cluster 0 (smaller cluster) - logarithmic scale

- "Experimental Support for a Categorical Compositional Distributional Model of Meaning" [89] – a paper about "modelling compositional meaning for sentences using empirical distributional methods".
- "Gaussian error linear units (gelus)" [90] – paper introducing GELU, a new activation function in neural networks, which was extensively tested in future research [160].

Each of these papers introduced novel, very innovative ideas that inspired further research directions. They can be thus treated as belonging to a unique (separate) subset of contributions.

3.8. RQ6: Text comprehension

Finally, an additional aspect of text belonging to the dataset was measured; text comprehensibility. This is a very complicated problem, which is still being explored. Taking into account that one of the considered audiences are researchers interested in starting work in NLP, text difficulty, using existing text complexity metrics, was evaluated. An important note is that these metrics are known for problems, such as: not considering complicated mathematical formula; skipping charts, pictures and other visuals. Keeping this in mind, let us proceed further.

3.8.1. Text complexity

The most common comprehensibility measures map text to school grade, in the American education system [28]. In this way, it is established what is the expected level of reader that should be able to understand the text. The used measures were:

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

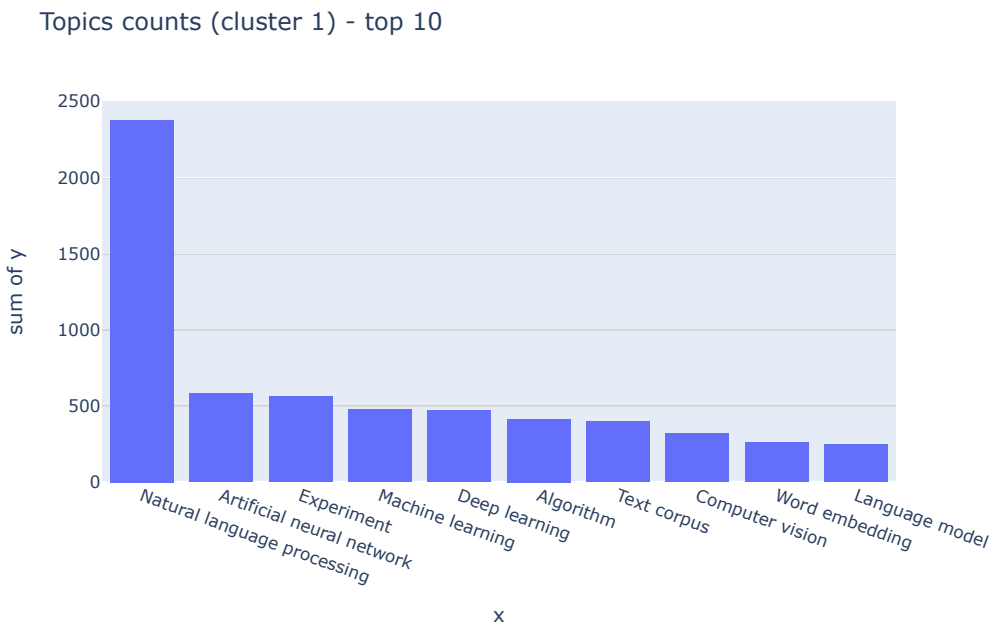

Figure 20: Histogram of topics counts in cluster 1 (bigger cluster)

- Flesch Reading Ease [23]
- Flesch Kincaid Grade [23]
- Gunning Fog [24]
- Smog Index [25]
- Automated Readability Index [23]
- Coleman Liau Index [26]
- Linsear Write Formula [27]

All measures return results on equal scale (school grade). Furthermore, they were all consistent in terms of paper scores. To provide the least biased results, the numerical values (Section 3.8.2) have been averaged to achieve a single, straightforward, measure for text complexity. Here, it should be noted that this was done also because delving into discussion of ultimate validity of individual comprehensibility measurements and pros/cons of each of them is out of scope of current contribution. Rather, the combined measure was calculated to obtain a general idea as to the "readability" of the literature in question.

The results can be averaged together between metrics, because all of they refer to the same scale (school grade).

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

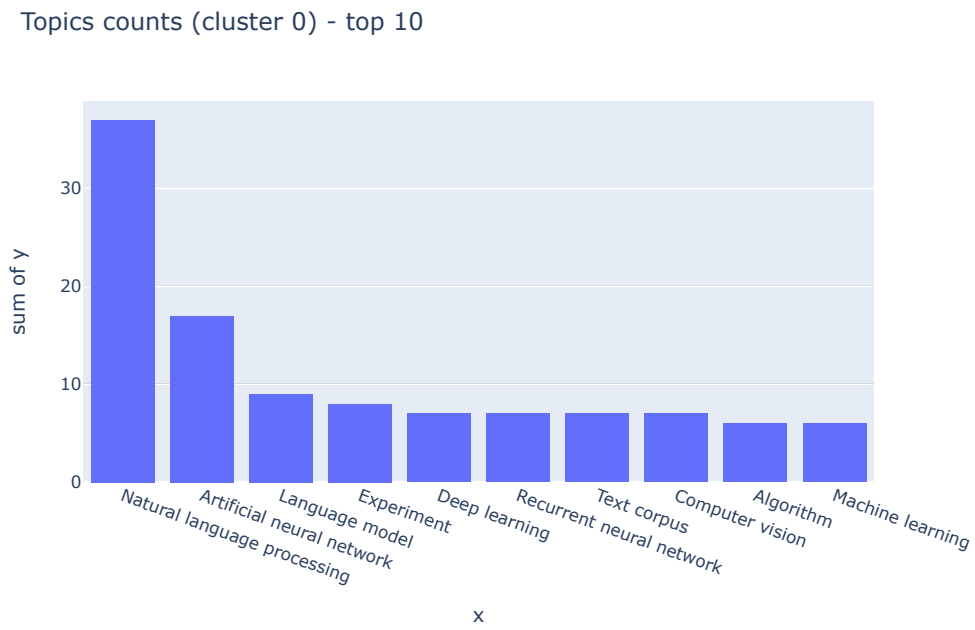

Figure 21: Histogram of topics counts in cluster 0 (smaller cluster)

3.8.2. RQ6: establishing complexity level of NLP literature

Results of the text complexity (RQ6) are rather intuitive.

As shown in Figure 24, the averaged score of 15 comprehensibility metrics suggests that the majority of papers, in the NLP domain, can be understood by a person after "15th grade". This matches roughly a person who finished the "1st stage" of college education (engineering studies, bachelor degree, and similar). Obviously, this result shows that use of such metrics to "scientific texts" has limited applicability, as they are based mostly on syntactic features of the text, while the semantics makes some of them difficult to follow even for the specialists. This, particularly, applies to texts which contain mathematical equations, which are being removed during text preprocessing.

3.9. Summary of key results

Let us now summarize the key finding, in the form of a question-answer for each of RQs that have been postulated in Section 1.

RQ1: What datasets are considered to be most useful?

The datasets used most commonly for NLP research are: Wikipedia, Twitter, Facebook, WordNet, arXiv, Academic, SST (The Stanford Sentiment Treebank), SQuAD (The Stanford Question Answering Dataset), NLI and SNLI (Stanford Natural Language Inference Corpus), COCO (Common Objects in Context), Reddit.

RQ2: Which languages, other than English, appear as a topic of NLP research?

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

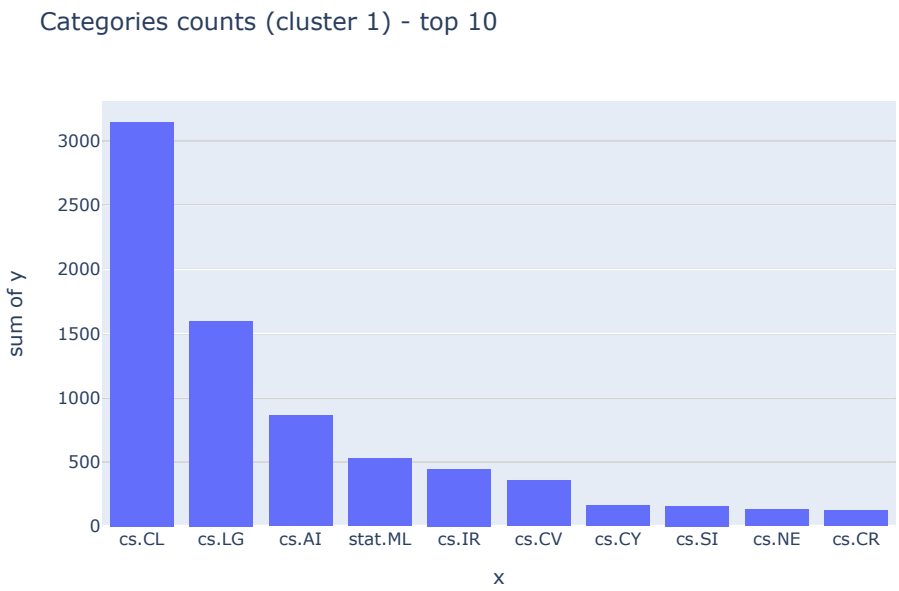

Figure 22: Histogram of categories counts in cluster 1 (bigger cluster)

Languages analyzed most commonly in NLP research, apart from English and Chinese, are: German, French and Spanish.

RQ3: What are the most popular fields and topics in current NLP research?

The most popular fields studied in NLP literature are: Natural Language Processing/Language Computing, artificial intelligence, machine learning, neural networks and deep learning and text embedding.

RQ4: What particular tasks and problems are most often studied?

Particular tasks and problems, which appear in the literature, are: text embedding with BERT and transformers, machine translation between English and other languages (especially English-Chinese), sentiment analysis (most popular with Twitter and Wikipedia datasets), question answering models (with Wikipedia and SQuAD datasets), named entity recognition, and text summarization.

RQ5: Is the field "homogenous", or are there easily identifiable "subgroups"?

According to the text embedding analysis, there is not enough evidence to find a strongly distinguishable clusters. Hence, there are no outstanding subgroups in the NLP literature.

RQ6: How difficult is it to comprehend the NLP literature?

According to averaged standard comprehensibility measures, scientific texts related to NLP can be digested by a 15th graders, which maps to the 3rd year of higher education (e.g. College, Bachelor's degree studies etc.)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

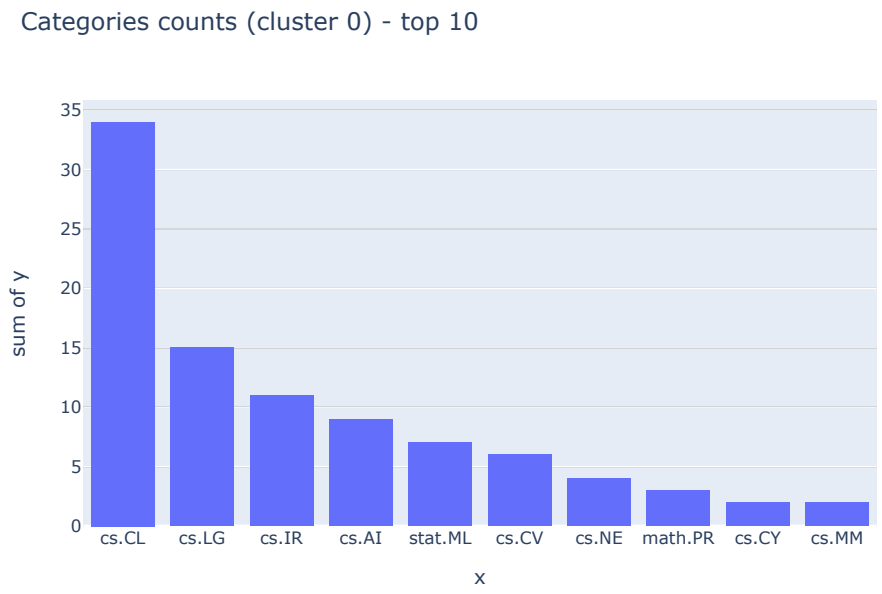

Figure 23: Histogram of categories counts in cluster 0 (smaller cluster)

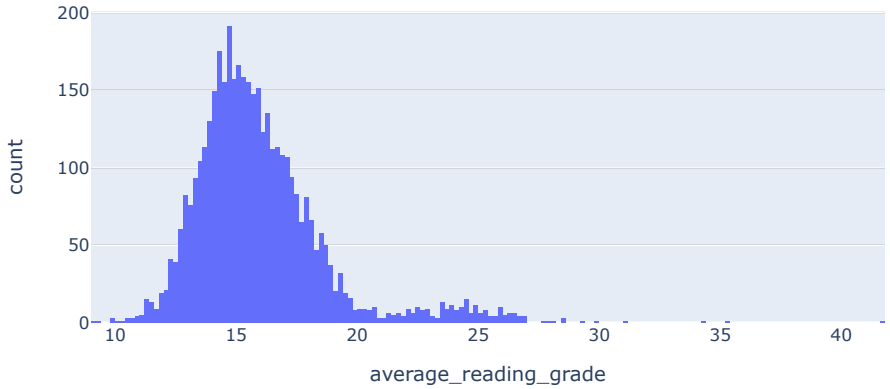

Figure 24: Average reading grade (mean of all metrics; bottom 99th percentile) histogram showing what grade should the reader be to understand the papers.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

4. Concluding remarks

This analysis used Natural Language Processing methods to analyze scientific literature related to NLP. The goal was to answer 6 research questions (RQ1-RQ6). A total of 4712 scientific papers in the field of NLP from arXiv were analyzed. The work used and illustrated at the same time the following NLP methods: text extraction, text cleaning, text preprocessing, keyword and keyphrase search, text embeddings, abstractive and extractive text summarization, text complexity and other methods such as: clustering, metadata analysis, citation/reference analysis, network visualization. This analysis focuses on only Natural Language Processing and its subdomains, topics, etc. Since the procedures of obtaining results reported here were fully automated, the same or similar analysis could be analogically done with ease for different literature languages and even fields. Hence, all the tools used for the analysis are available in a designated repository [32] for future applications.

References

[1] Zhang, Chi, Shagan Sah, Thang Nguyen, Dheeraj Peri, Alexander Loui, Carl Salvaggio, and Raymond Ptucha. ”Semantic sentence embeddings for paraphrasing and text summarization.” In 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 705-709. IEEE, 2017.

[2] Ponti, Edoardo Maria, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. ”Modeling language variation and universals: A survey on typological linguistics for natural language processing.” Computational Linguistics 45, no. 3 (2019): 559-601.

[3] Kong, Yuqi, Fanchao Meng, and Benjamin Carterette. ”A Topological Method for Comparing Document Semantics.” arXiv preprint arXiv:2012.04203 (2020).

[4] Siegel, Eric V. ”Learning methods for combining linguistic indicators to classify verbs.” arXiv preprint cmp-lg/9709002 (1997).

[5] Font, Joel Escudé, and Marta R. Costa-Jussa. ”Equalizing gender biases in neural machine translation with word embeddings techniques.” arXiv preprint arXiv:1901.03116 (2019).

[6] Mayr, Philipp, Muthu Kumar Chandrasekaran, and Kokil Jaidka.
”Report on the 3rd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (birndl 2018).” In ACM SIGIR Forum, vol. 52, no. 2, pp. 105-110. New York, NY, USA: ACM, 2019.

[7] Kotson, Michael C., and Alexia Schulz. ”Characterizing phishing threats with natural language processing.” In 2015 IEEE Conference on Communications and Network Security (CNS), pp. 308-316. IEEE, 2015.

[8] Audenaert, Koenraad MR. ”Quantum skew divergence.” Journal of Mathematical Physics 55, no. 11 (2014): 112202.

[9] Zeng, William, and Bob Coecke. ”Quantum algorithms for compositional natural language processing.” arXiv preprint arXiv:1608.01406 (2016).

[10] Lu, Kaiji, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. ”Gender bias in neural natural language processing.” In Logic, Language, and Security, pp. 189-202. Springer, Cham, 2020.

[11] Thorne, Camilo, and Saber Akhondi. ”Word Embeddings for Chemical Patent Natural Language Processing.” arXiv preprint arXiv:2010.12912 (2020).

[12] Kajiwara, Tomoyuki, Chihiro Tanikawa, Yuujin Shimizu, Chenhui Chu, Takashi Yamashiro, and Hajime Nagahara. ”Using Natural Language Processing to Develop an Automated Orthodontic Diagnostic System.” arXiv preprint arXiv:1905.13601 (2019).

[13] Jensenius, Francesca R., Mala Htun, David J. Samuels, David A. Singer, Adria Lawrence, and Michael Chwe. “The Benefits and Pitfalls of Google Scholar.” PS: Political Science & Politics 51, no. 4 (2018): 820–24. doi:10.1017/S1049096518000940X.

[14] Gray, J.E. & Hamilton, M.C. & Hauser, A. & Janz, Margaret & Peters, J.P. & Taggart, Fiona. (2012). Scholarish: Google scholar and its value to the sciences. Issues in Science and Technology Librarianship. 70. 10.5062/F4MK69T9.

[32] `https://anonymous.4open.science/r/nlp-review-F81D`

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

[15] Sawicki, Jan, Maria Ganzha, Marcin Paprzycki, and Amelia Badica. "Exploring Usability of Reddit in Data Science and Knowledge Processing." Scalable Computing: Practice and Experience 23, no. 1 (2022): 9-22.
[16] Cielen, Davy, and Arno Meysman. Introducing data science: big data, machine learning, and more, using Python tools. Simon and Schuster, 2016.
[17] Paffenroth, Randy, and Xiangnan Kong. "Python in Data Science Research and Education." In Proceedings of the 14th Python in Science Conference, doi: 10.25080/Majora-7b98e3ed, vol. 19, pp. 164-170. 2015.
[18] Nagpal, Abhinav, and Goldie Gabrani. "Python for data analytics, scientific and technical applications." In 2019 Amity international conference on artificial intelligence (AICAI), pp. 140-145. IEEE, 2019.
[19] Soumya, K., G. Ramanathan, and G. Clinton. "An Assessment on Classification in Python Using Data Science." In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 551-555. IEEE, 2021.
[20] Şahinaslan, Ender. "Review of the most popular data science programs used today: python and r." (2019).
[21] Srinath, K. R. "Python–the fastest growing programming language." International Research Journal of Engineering and Technology (IRJET) 4, no. 12 (2017): 354-357.
[22] Li Jun and Li Ling,
"Comparative research on Python speed optimization strategies," 2010 International Conference on Intelligent Computing and Integrated Systems, 2010, pp. 57-59, doi: 10.1109/ICISS.2010.5655011.
[23] Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch, 1975.
[24] Gunning, Robert. "The fog index after twenty years." Journal of Business Communication 6, no. 2 (1969): 3-13.
[25] Mc Laughlin, G. Harry. "SMOG grading-a new readability formula." Journal of reading 12, no. 8 (1969): 639-646.
[26] Coleman, Meri, and Ta Lin Liau. "A computer readability formula designed for machine scoring." Journal of Applied Psychology 60, no. 2 (1975): 283.
[27] Eltorai, Adam EM, Syed S. Naqvi, Soha Ghanian, Craig P. Eberson, Arnold-Peter C. Weiss, Christopher T. Born, and Alan H. Daniels. "Readability of invasive procedure consent forms." Clinical and translational science 8, no. 6 (2015): 830-833.
[28] Spring, Joel. American education. Routledge, 2019.
[29] Lin, Jiahuang, Xin Li, and Gennady Pekhimenko. "Multi-node Bert-pretraining: Cost-efficient approach." arXiv preprint arXiv:2008.00177 (2020).
[30] Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. "Advances in pre-training distributed word representations." arXiv preprint arXiv:1712.09405 (2017).
[31] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[32] Zaharia, Matei, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker, and Ion Stoica. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." In 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), pp. 15-28. 2012.
[33] De Bortoli Fávero, Eliane M., Dalcimar Casanova, and Andrey Ricardo Pimentel. "SE3M: A Model for Software Effort Estimation Using Pre-trained Embedding Models." arXiv e-prints (2020): arXiv-2006.
[34] Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. "FLAIR: An easy-to-use framework for state-of-the-art NLP." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 54-59. 2019.
[35] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.
[36] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
[37] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In Advances in neural information processing systems, pp. 3111-3119. 2013.
[38] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." OpenAI blog 1, no. 8 (2019): 9.
[39] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
[40] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683 (2019).

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

[41] HDBSCAN Documentation, Benchmarking Performance and Scaling of Python Clustering Algorithms, accessed 13 Nov 2021, https:hdbscan.readthedocs.ioen0.8.18performance_and_scalability.html
[42] Xu, Dongkuan, and Yingjie Tian. "A comprehensive survey of clustering algorithms." Annals of Data Science 2, no. 2 (2015): 165-193.
[43] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. 11 (2008).
[44] Claassen, Christopher. "Improving t-SNE for applications on word embedding data in text mining." (2020).
[45] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, pp. 281-297. 1967.
[46] Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. "OPTICS: Ordering points to identify the clustering structure." ACM Sigmod record 28, no. 2 (1999): 49-60.
[47] Schubert, Erich, and Michael Gertz. "Improving the Cluster Structure Extracted from OPTICS Plots." In LWDA, pp. 318-329. 2018.
[48] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In kdd, vol. 96, no. 34, pp. 226-231. 1996.
[49] Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN." ACM Transactions on Database Systems (TODS) 42, no. 3 (2017): 1-21.
[50] Farhadur Rahman, Md, Weimo Liu, Saad Bin Suhaim, Saravanan Thirumuruganathan, Nan Zhang, and Gautam Das. "HDBSCAN: Density based Clustering over Location Based Services." arXiv e-prints (2016): arXiv-1602.
[51] McInnes, Leland, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." Journal of Open Source Software 2, no. 11 (2017): 205.
[52] Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." ACM sigmod record 25, no. 2 (1996): 103-114.
[53] Liashchynskyi, Petro, and Pavlo Liashchynskyi. "Grid search, random search, genetic algorithm: a big comparison for NAS." arXiv preprint arXiv:1912.06059 (2019).
[54] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.
[55] Davies, David L., and Donald W. Bouldin. "A cluster separation measure." IEEE transactions on pattern analysis and machine intelligence 2 (1979): 224-227.
[56] Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." Communications in Statistics-theory and Methods 3, no. 1 (1974): 1-27.
[57] Dalal, Vipul, and Latesh Malik. "A survey of extractive and abstractive text summarization techniques." In 2013 6th International Conference on Emerging Trends in Engineering and Technology, pp. 109-110. IEEE, 2013.
[58] Bhatia, Neelima, and Arunima Jaiswal. "Trends in extractive and abstractive techniques in text summarization." International Journal of Computer Applications 117, no. 6 (2015).
[59] Miller, Derek. "Leveraging BERT for extractive text summarization on lectures." arXiv preprint arXiv:1906.04165 (2019).
[60] Luhn, Hans Peter. "The automatic creation of literature abstracts." IBM Journal of research and development 2, no. 2 (1958): 159-165.
[61] Steinberger, Josef, and Karel Jezek. "Using latent semantic analysis in text summarization and summary evaluation." Proc. ISIM 4 (2004): 93-100.
[62] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." In Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404-411. 2004.
[63] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." Journal of artificial intelligence research 22 (2004): 457-479.
[64] Haghighi, Aria, and Lucy Vanderwende. "Exploring content models for multi-document summarization." In Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics, pp. 362-370. 2009.
[65] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).
[66] Gidiotis, Alexios, and Grigorios Tsoumakas. "A divide-and-conquer approach to the summarization of long documents." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 3029-3040.
[67] Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).
[68] Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In International Conference on Machine Learning, pp. 11328-11339. PMLR, 2020.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

[69] Azzam, Tarek, Stephanie Evergreen, Amy A. Germuth, and Susan J. Kistler. "Data visualization and evaluation." New Directions for Evaluation 2013, no. 139 (2013): 7-32.
[70] Pousman, Zachary, John Stasko, and Michael Mateas. "Casual information visualization: Depictions of data in everyday life." IEEE transactions on visualization and computer graphics 13, no. 6 (2007): 1145-1152.
[71] Card, Mackinlay. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.
[72] Tufte, Edward. "The visual display of quantitative information." (2001).
[73] Cairo, Alberto. The Functional Art: An introduction to information graphics and visualization. New Riders, 2012.
[74] American Statistical Association. "Publications of the American Statistical Association." American Statistical Association, 1914.
[75] Bertini, E., N. Elmqvist, and T. Wischgoll. "Judgment error in pie chart variations." In Proceedings of the Eurographics/IEEE VGTC conference on visualization: Short papers, pp. 91-95. 2016.
[76] Kosara, Robert. "Evidence for area as the primary visual cue in pie charts." In 2019 IEEE Visualization Conference (VIS), pp. 101-105. IEEE, 2019.
[77] Goldberg, Joseph, and Jonathan Helfman. "Eye tracking for visualization evaluation: Reading values on linear versus radial graphs." Information visualization 10, no. 3 (2011): 182-195.
[78] Wheildon, Colin, and Geoffrey Heard. Type & layout: Are you communicating or just making pretty shapes. Worsley Press, 2005.
[79] Stone, Maureen. "Choosing colors for data visualization." Business Intelligence Network 2 (2006).
[80] Tufte, Edward R., Nora Hillman Goeler, and Richard Benson. Envisioning information. Vol. 2. Cheshire, CT: Graphics press, 1990.
[81] Kirschbaum, Charles. "Network analysis: emergence, criticism and recent trends." RAUSP Management Journal 54 (2019): 533-547.
[82] Csárdi, Gábor, and Tamás Nepusz. "igraph Reference manual." URL: http://igraph. sourceforge. net/documentation. html (accessed April 20 (2010).
[83] Hagberg, Aric, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
[84] Baird, Laura M., and Charles Oppenheim. "Do citations matter?." Journal of information Science 20, no. 1 (1994): 2-15.
[85] Aksnes, Dag W., Liv Langfeldt, and Paul Wouters. "Citations, citation indicators, and research quality: An overview of basic concepts and theories." Sage Open 9, no. 1 (2019): 2158244019829575.
[86] Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467 (2016).
[87] Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).
[88] Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. "Natural language processing (almost) from scratch." Journal of machine learning research 12, no. ARTICLE (2011): 2493-2537.
[89] Grefenstette, Edward, and Mehrnoosh Sadrzadeh. "Experimental support for a categorical compositional distributional model of meaning." arXiv preprint arXiv:1106.4058 (2011).
[90] Hendrycks, Dan, and Kevin Gimpel. "Gaussian error linear units (gelus)." arXiv preprint arXiv:1606.08415 (2016).
[91] Zhang, Qiao, Cong Wang, Hongyi Wu, Chunsheng Xin, and Tran V. Phuong. "GELU-Net: A Globally Encrypted, Locally Unencrypted Deep Neural Network for Privacy-Preserved Learning." In IJCAI, pp. 3933-3939. 2018.
[92] Hendrycks, Dan, and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks." arXiv preprint arXiv:1610.02136 (2016).
[93] Tsuchida, Russell, Tim Pearce, Chris van der Heide, Fred Roosta, and Marcus Gallagher. "Avoiding Kernel Fixed Points: Computing with ELU and GELU Infinite Networks." arXiv preprint arXiv:2002.08517 (2020).
[94] Fellows, Ian Edward. Exponential family random network models. University of California, Los Angeles, 2012.
[95] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164. 2015.
[96] Wu, Qi, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. "Visual question answering: A survey of methods and datasets." Computer Vision and Image Understanding 163 (2017): 21-40.
[97] Liu, Chenxi, Junhua Mao, Fei Sha, and Alan Yuille. "Attention correctness in neural image captioning." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1. 2017.
[98] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

preprint arXiv:2010.11929 (2020).
[99] Peters, M. E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, and Kenton Lee. "Deep contextualized word representations. arXiv 1802.05365. doi: 10.18653/v1." N18-1202 (2018).
[100] Honnibal, M., and I. Montani. "Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." Unpublished software application. https://spacy. io (2017).
[101] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019).
[102] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019).
[103] Meng, Chen, Minmin Sun, Jun Yang, Minghui Qiu, and Yang Gu. "Training deeper models by GPU memory optimization on TensorFlow." In Proc. of ML Systems Workshop in NIPS. 2017.
[104] Fellbaum, Christiane. "WordNet: An electronic lexical resource." The Oxford handbook of cognitive science (2017): 301-314.
[105] Siddhant, Aditya, and Zachary C. Lipton. "Deep bayesian active learning for natural language processing: Results of a large-scale empirical study." arXiv preprint arXiv:1808.05697 (2018).
[106] Jang, Beakcheol, Inhwan Kim, and Jong Wook Kim. "Word2vec convolutional neural networks for classification of news articles and tweets." PloS one 14, no. 8 (2019): e0220976.
[107] Kim, Jin-Dong, Sang-Zoo Lee, and Hae-Chang Rim. "HMM specialization with selective lexicalization." arXiv preprint cs/9912016 (1999).
[108] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. Ieee, 2009.
[109] Hongtao, Lu, and Zhang Qinchuan. "Applications of deep convolutional neural network in computer vision." Journal of Data Acquisition and Processing 31, no. 1 (2016): 1-17.
[110] Hu, Yu Hen, and Jeng-Neng Hwang, eds. "Handbook of neural network signal processing." (2002): 2525-2526.
[111] Abiodun, Oludare Isaac, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. "State-of-the-art in artificial neural network applications: A survey." Heliyon 4, no. 11 (2018): e00938.
[112] Kwon, Donghwoon, Hyunjoo Kim, Jinoh Kim, Sang C. Suh, Ikkyun Kim, and Kuinam J. Kim. "A survey of deep learning-based network anomaly detection." Cluster Computing 22, no. 1 (2019): 949-961.
[113] Baxt, William G. "Application of artificial neural networks to clinical medicine." The lancet 346, no. 8983 (1995): 1135-1138.
[114] Iftene, Adrian. "Exploiting Social Networks. Technological Trends (Habilitation Thesis)." arXiv preprint arXiv:2004.14386 (2020).
[115] Sharifirad, Sima, Borna Jafarpour, and Stan Matwin. "How is your mood when writing sexist tweets? detecting the emotion type and intensity of emotion using natural language processing techniques." arXiv preprint arXiv:1902.03089 (2019).
[116] Zubiaga, Arkaitz. "Mining social media for newsgathering: A review." Online Social Networks and Media 13 (2019): 100049.
[117] Wankmüller, Sandra. "Neural Transfer Learning with Transformers for Social Science Text Analysis." arXiv preprint arXiv:2102.02111 (2021).
[118] Sarkar, Kamal. "JU_KS@ SAIL_CodeMixed-2017: Sentiment Analysis for Indian Code Mixed Social Media Texts." arXiv preprint arXiv:1802.05737 (2018).
[119] Amano, Tatsuya, Juan P. González-Varo, and William J. Sutherland. "Languages are still a major barrier to global science." PLoS biology 14, no. 12 (2016): e2000933.
[120] Fishman, Joshua A., and Charles Albert Ferguson. Language problems of developing nations. Edited by Jyotirindra Dasgupta. New York: Wiley, 1968.
[121] Alousque, Isabel Negro. "Cultural domains: Translation problems." Revista de lingüística y lenguas aplicadas 4, no. 1 (2009): 137-145.
[122] SCHWARZ, Narcisa, Laura-Rebeca STIEGELBAUER, and Diana-Bianca HUSAR. "TRANSLATION PROBLEMS AND DIFFICULTIES IN APPLIED TRANSLATION PROCESSES." Studii De Ştiintă Şi Cultură 12, no. 3 (2016).
[123] Akan, Md Faruquzzaman, Md Rezaul Karim, and Abdullah Mohammad Kabir Chowdhury. "An analysis of Arabic-English translation: Problems and prospects." Advances in Language and Literary Studies 10, no. 1 (2019): 58-65.
[124] Brazill, Shihua Chen. "Chinese to English translation: Identifying problems and providing solutions." PhD diss., Montana Tech of The University of Montana, 2016.
[125] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT:

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).
[126] Pearce, Kate, Tiffany Zhan, Aneesh Komanduri, and Justin Zhan. "A Comparative Study of Transformer-Based Language Models on Extractive Question Answering." arXiv preprint arXiv:2110.03142 (2021).
[127] Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained models for natural language processing: A survey." Science China Technological Sciences (2020): 1-26.
[128] Rietzler, Alexander, Sebastian Stabinger, Paul Opitz, and Stefan Engl. "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification." arXiv preprint arXiv:1908.11860 (2019).
[129] Rusnachenko, Nicolay, and Natalia Loukachevitch. "Studying Attention Models in Sentiment Attitude Extraction Task." In International Conference on Applications of Natural Language to Information Systems, pp. 157-169. Springer, Cham, 2020.
[130] DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. "Eraser: A benchmark to evaluate rationalized nlp models." arXiv preprint arXiv:1911.03429 (2019).
[131] Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond accuracy: Behavioral testing of NLP models with CheckList." arXiv preprint arXiv:2005.04118 (2020).
[132] Bragg, Jonathan, Arman Cohan, Kyle Lo, and Iz Beltagy. "Flex: Unifying evaluation for few-shot nlp." In Thirty-Fifth Conference on Neural Information Processing Systems. 2021.
[133] Rodriguez, Pedro, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. "Evaluation Examples Are Not Equally Informative: How Should That Change NLP Leaderboards?." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4486-4503. 2021.
[134] Clark, Stephen, Bob Coecke, Edward Grefenstette, Stephen Pulman, and Mehrnoosh Sadrzadeh. "A quantum teleportation inspired algorithm produces sentence meaning from word meaning and grammatical structure." arXiv preprint arXiv:1305.0556 (2013).
[135] O'Riordan, Lee J., Myles Doyle, Fabio Baruffa, and Venkatesh Kannan. "A hybrid classical-quantum workflow for natural language processing." Machine Learning: Science and Technology 2, no. 1 (2020): 015011.
[136] Ishtiaq, Arhum, and Sara Mahmood. "Quantum Machine Learning: Fad or Future?." arXiv preprint arXiv:2106.10714 (2021).
[137] Kartsaklis, Dimitri, Ian Fan, Richie Yeung, Anna Pearson, Robin Lorenz, Alexis Toumi, Giovanni de Felice, Konstantinos Meichanetzidis, Stephen Clark, and Bob Coecke. "lambeq: An Efficient High-Level Python Library for Quantum NLP." arXiv preprint arXiv:2110.04236 (2021).
[138] Meichanetzidis, Konstantinos, Stefano Gogioso, Giovanni De Felice, Nicolò Chiappori, Alexis Toumi, and Bob Coecke. "Quantum natural language processing on near-term quantum computers." arXiv preprint arXiv:2005.04147 (2020).
[139] Liu, Guanxiong, Issa Khalil, and Abdallah Khreishah. "GanDef: A GAN based adversarial training defense for neural network classifier." In IFIP International Conference on ICT Systems Security and Privacy Protection, pp. 19-32. Springer, Cham, 2019.
[140] Juba, Brendan, and Hai S. Le. "Precision-recall versus accuracy and the role of large data sets." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 4039-4048. 2019.
[141] Tian, Fei, Bin Gao, En-Hong Chen, and Tie-Yan Liu. "Learning better word embedding by asymmetric low-rank projection of knowledge graph." Journal of Computer Science and Technology 31, no. 3 (2016): 624-634.
[142] Antognini, Diego, and Boi Faltings. "Learning to create sentence semantic relation graphs for multi-document summarization." arXiv preprint arXiv:1909.12231 (2019).
[143] Koner, Rajat, Hang Li, Marcel Hildebrandt, Deepan Das, Volker Tresp, and Stephan Günnemann. "Graphhopper: Multi-hop Scene Graph Reasoning for Visual Question Answering." In International Semantic Web Conference, pp. 111-127. Springer, Cham, 2021.
[144] Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. "Carbon emissions and large neural network training." arXiv preprint arXiv:2104.10350 (2021).
[145] Leroy, Suzanne AG. "From natural hazard to environmental catastrophe: Past and present." Quaternary International 158, no. 1 (2006): 4-12.
[146] Scheffer, Marten, Steve Carpenter, Jonathan A. Foley, Carl Folke, and Brian Walker. "Catastrophic shifts in ecosystems." Nature 413, no. 6856 (2001): 591-596.
[147] Heing, Bridey, ed. Environmental Catastrophe. Greenhaven Publishing LLC, 2019.
[148] Wiedmann, Thomas. "Carbon footprint and input–output analysis–an introduction." (2009): 175-186.
[149] Sippel, Sebastian, Nicolai Meinshausen, Erich M. Fischer, Enikő Székely, and Reto Knutti. "Climate change now detectable from any single day of weather at global scale." Nature Climate Change 10, no. 1 (2020): 35-41.
[150] AAbraham de Moivre, "Approximatio ad summam terminorum binomii a+bn in seriem expansi." (1738)
[151] Walker, Helen Mary. "DE MOIVRE ON THE LAW OF NORMAL PROBABILITY." (2006).

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

[152] Marsaglia, George. ”Evaluating the normal distribution.” Journal of Statistical Software 11, no. 4 (2004): 1-7.
[153] Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. ”Improving language understanding by generative pre-training.” (2018).
[154] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. ”Language models are few-shot learners.” arXiv preprint arXiv:2005.14165 (2020).
[155] Xu, Anbang, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. ”A new chatbot for customer service on social media.” In Proceedings of the 2017 CHI conference on human factors in computing systems, pp. 3506-3510. 2017.
[156] Höchstötter, Nadine, and Dirk Lewandowski. ”What users see–Structures in search engine results pages.” Information Sciences 179, no. 12 (2009): 1796-1812.
[157] Handoyo, Eko, M. Arfan, Yosua Alvin Adi Soetrisno, Maman Somantri, Aghus Sofwan, and Enda Wista Sinuraya. ”Ticketing chatbot service using serverless NLP technology.” In 2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), pp. 325-330. IEEE, 2018.
[158] Hohenstein, Jess, Dominic DiFranzo, Rene F. Kizilcec, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeff Hancock, and Malte Jung. ”Artificial intelligence in communication impacts language and social relationships.” arXiv preprint arXiv:2102.05756 (2021).
[159] Rehurek, Radim, and Petr Sojka. ”Gensim–python framework for vector space modelling.” NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3, no. 2 (2011).
[160] Huang, Zhiguan, Xiaohao Du, Liangming Chen, Yuhe Li, Mei Liu, Yao Chou, and Long Jin. ”Convolutional neural network based on complex networks for brain tumor image classification with a modified activation function.” IEEE Access 8 (2020): 89281-89290.
[161] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. ”Attention is all you need.” In Advances in neural information processing systems, pp. 5998-6008. 2017.
[162] Sun, Lu, Mingtian Tan, and Zhe Zhou. ”A survey of practical adversarial example attacks.” Cybersecurity 1, no. 1 (2018): 1-9.
[163] Floridi, Luciano, and Massimo Chiriatti. ”GPT-3: Its nature, scope, limits, and consequences.” Minds and Machines 30, no. 4 (2020): 681-694.
[164] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. ”Energy and policy considerations for deep learning in NLP.” arXiv preprint arXiv:1906.02243 (2019).
[165] Ahmet, Ahmed, and Tariq Abdullah. ”Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach.” In 2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE), pp. 41-48. IEEE, 2020.
[166] Gupta, Tanya, and Gurukul Vidyapeeth. ”Keyword extraction: a review.” International Journal of Engineering Applied Sciences and Technology 2, no. 4 (2017): 215-220.
[167] Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. ”Understanding bag-of-words model: a statistical framework.” International Journal of Machine Learning and Cybernetics 1, no. 1 (2010): 43-52.
[168] Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. ”Enriching word vectors with subword information.” Transactions of the association for computational linguistics 5 (2017): 135-146.
[169] Voss, Jakob. ”Measuring wikipedia.” (2005).
[170] Cha, Miriam, Youngjune Gwon, and H. T. Kung. ”Language modeling by clustering with word embeddings for text readability assessment.” In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2003-2006. 2017.
[171] Millman, K. Jarrod, and Michael Aivazis. ”Python for scientists and engineers.” Computing in Science & Engineering 13, no. 2 (2011): 9-12.
[172] Liu, Honghua, Jing Yang, Ming Ye, Scott C. James, Zhonghua Tang, Jie Dong, and Tongju Xing. ”Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data.” Journal of Hydrology 597 (2021): 126146.
[173] Liu, Shusen, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. ”Visual exploration of semantic relationships in neural word embeddings.” IEEE transactions on visualization and computer graphics 24, no. 1 (2017): 553-562.
[174] Hinton, Geoffrey E., and Sam Roweis. ”Stochastic neighbor embedding.” Advances in neural information processing systems 15 (2002).
[175] Yanova, Natalia. ”The intelligent systems for language phonetic interfaces.” In Proceedings of the 14th IADIS International Conference Computer Graphics, Visualization, Computer Vision and Image Processing 2020, MCCSIS 2020, pp. 257-261. 2020.
[176] Konieczny, Jakub. T̈raining of neural machine translation model to apply terminology constraints for language with robust inflection.Ännals of Computer Science and Information Systems 26 (2021): 233234.
[177] Jarmosz, Wojciech. ”Applying Machine Translation Methods in the Problem of Automatic Text Correction.” Posi-

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00213

tion and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems (September 26, 2021). doi:10.15439/2021f142.

[178] Hanslo, Ridewaan. "Evaluation of Neural Network Transformer Models for Named-Entity Recognition on Low-Resourced Languages." In 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 115-119. IEEE, 2021.

[179] Goyal, Archana, Vishal Gupta, and Manish Kumar. "Recent named entity recognition and classification techniques: a systematic review." Computer Science Review 29 (2018): 21-43.

[180] Mohit, Behrang. "Named entity recognition." In Natural language processing of semitic languages, pp. 221-245. Springer, Berlin, Heidelberg, 2014.

[181] Siddiqi, Sifatullah, and Aditi Sharan. "Keyword and keyphrase extraction techniques: a literature review." International Journal of Computer Applications 109, no. 2 (2015).