# A Bibliometric Review of Methods and Algorithms for Generating Corpora for Learning Vector Word Embeddings

Beibarys Sagingaliyev[1], Zhuldyzay Aitakhunova[1], Adel Shaimerdenova[1],
Iskander Akhmetov[1,2](✉), Alexander Pak[1,2], and Assel Jaxylykova[2,3]

[1] Kazakh-British Technical University, Almaty, Kazakhstan
{b_sagingaliyev,z_aitakhunova,a_shaimerdenova}@kbtu.kz
[2] Institute of Information and Computational Technologies, Almaty, Kazakhstan
i.akhmetov@ipic.kz
[3] Al-Farabi Kazakh National University, Almaty, Kazakhstan

**Abstract.** Natural Language Processing (NLP) problems are among the hardest Machine Learning (ML) problems due to the complex nature of the human language. The introduction of word embeddings improved the performance of ML models on various NLP tasks as text classification, sentiment analysis, machine translation, etc. Word embeddings are real-valued vector representations of words in a specific vector space. Producing quality word embeddings that are then used as input to downstream NLP tasks is important in obtaining a good performance. To accomplish it, corpora of sufficient size is needed. Corpora may be formed in a multitude of ways, including text that was originally electronic, spoken language transcripts, optical character recognition, and synthetically producing text from the available dataset. The study provides the most recent bibliometric analysis on the topic of corpora generation for learning word vector embeddings. The analysis is based on the publication data from 2006 to 2022 retrieved from Scopus scientific database. A descriptive analysis method has been employed to obtain statistical characteristics of the publications in the research area. The systematic analysis results show the field's evolution over time and highlight influential contributions to the field. It is believed that compiled bibliometric reviews could help researchers gain knowledge of the general state of the scientific knowledge, its descriptive features, patterns, and insights to design their studies systematically.

**Keywords:** Corpora · Data augmentation · Word embedding · NLP downstream tasks · Descriptive analysis · Natural language processing

NLP word embedding models can now preserve semantic and syntactic features in generated data from large collections of unlabeled texts known as corpora. The collection of rules used to analyze a language can choose and process it. Corpora may be formed in various ways, including text that was originally electronic, transcripts of spoken language, and optical character recognition. There

O. Pichardo Lagunas et al. (Eds.): MICAI 2022, LNAI 13613, pp. 148–162, 2022.
https://doi.org/10.1007/978-3-031-19496-2_12

are previously familiar methods, such as Word2vec, GloVe, and fastText, which improved the NLP tasks to upgrade. The latest modified models are Bert, Gpt-3 which are relatively new for the industry, developed by Google in 2018. This bibliometric review will illustrate investigating models in word embedding methods with data augmentations on NLP tasks.

## 1 Introduction

Word embedding is a term in Natural Language Processing, which refers to the language modeling and representing the words and whole sentences in a vector format [17]. There has been an ultimate rise in several downstream tasks such as question answering, text classifications, and sentiment analysis [12–14]. Nowadays, word embedding models can preserve semantic and syntactic features in generated data from an enormous unlabeled collection of texts called corpora. It is selected and processed by the set of rules leveraged to study a language [1]. The term corpora can be created in various methods, including text that was initially electronic, transcripts of spoken language, and other optical character recognition [3,15].

Bibliometrics is an analysis of books, articles and other publications using statistical methods, emphasizing scientific content in terms of objectivity and quantity. Bibliometric methods are widely used in librarianship, and informatics [4]. There are several outlining advantages of bibliometrics. First is organizing specific data from related works [10]. The second advantage is that it requires very little time to make and be used [9]. Thirdly, it is adaptable. It can be revised individually in institutional, national, and worldwide levels. Finally, the approaches are easy to deal with because they are based on simple counting. Because their use may be automated, several approaches have become extremely simple in the digital age [11].

Bibliometric analysis has been widely used in various fields to assess the quality and productivity of academic results, and it has proven to be highly effective over time. Relevant studies primarily focus on revealing statistical aspects of publications, discovering and exploring the collaborative relationship [6].

As a result, this study aims to analyze works related to generating corpora for learning vector word embeddings using bibliometric methodologies. To be more explicit, bibliometric data retrieved from Scopus, with the area of search in learning vector word embeddings downloaded and analyzed via descriptive analysis [2].

## 2 Methodology

We will analyze the bibliometrics by the following methods to get more insight into research related to word embedding methods with data augmentations on NLP tasks. Data retrieved from Scopus, with the search area in learning vector word embeddings.

### 2.1 Methods

*Descriptive analysis.* Descriptive analysis is the statistical method of statistical analysis by describing certain coefficients, which gives a brief explanation of the data. It could be a more general analysis or specific. Statistical measures derived from the descriptive analysis include standard deviation, mode, mean, skewness of the curve, etc. And variables are depicted in terms of graphs, charts, or tables, which gives a more visually informative understanding of statistical measures. Thus, it will help us gather knowledge on published research related to word embeddings and data augmentation tasks. It aims to generalize the large data set and give insightful features.

Here, we want to employ descriptive analysis to acquire statistically proven information on publications, citations, and authors. Also, their inter-relations and dependencies will be key points to analyze in this study. By analysis, we want to get the distribution of publications, citations, and average citations each year. Additionally, we will analyze publications by country, area of research, and journal. Thus, the objective is to get a clear analytical description of publications related to our research area.

### 2.2 Materials

Scopus is the main database for collecting the necessary information on publications related to corpora generation for learning word vector embeddings (the paper's topic). Scopus is considered a highly reputed and easily navigated multidisciplinary Elsevier citation database with high-quality peer-review and a large research base.

2050 rows of data, including publication authors, title, keywords, abstract, year, source journal, citation amount, etc., were retrieved from the database using "learning vector word embedding", "word embedding" and "text augmentation" keywords. The publication year of collected papers is between 2006 and 2022.

A list of publication information was downloaded in a CSV file format and analyzed in Python programming language using libraries such as pandas, NumPy, Matplotlib.

The descriptive analysis presented in the paper is based on information on the aforementioned 2050 publications. Figure 1 demonstrates publication distribution by subject areas: Engineering, Mathematics, Social Sciences, Decision Science, Medicine, Computer Science and etc. The top 3 areas are Computer Science, which contributes 43.6% of total publications, Engineering (13,3%), and Mathematics (10.7%). Table 1 and Figs. 2, 3, 4 and 5 present a statistical overview of collected publications. Results show that a total of 2050 publications are related with 2148 unique affiliations and 4894 unique authors. The average number of citations per publication is 11.95, the average number of authors is 3.51, and the average number of references is 33.46. An expanded analysis on the subject area, country/region, and affiliation can be found in the next section of the paper.

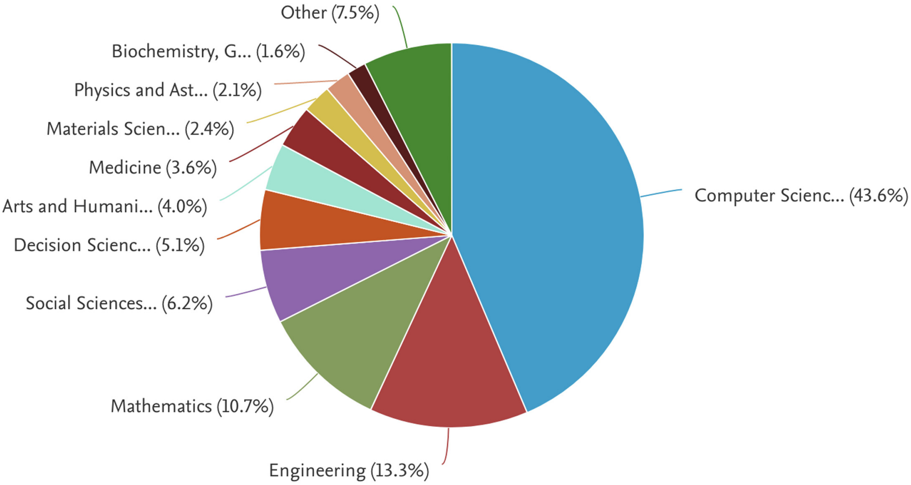

**Fig. 1.** Publication distribution by subject areas.

# 3 Results

## 3.1 Publications and Citations by Year

A descriptive analysis of the distribution of publication and citation by year is demonstrated in Fig. 6. The analysis results depict a bell-shaped distribution for total citation and a steadily increasing trend for total publications on word embedding and text augmentation research.

A total citation shows an upward increase until 2017 and then a gradual decline up to 2022. It can be explained by the fact that newer publications had less time to generate citations than older ones. Overall, it demonstrates that citations and interest by researchers in the area of NLP are increasing if we compare the total number of citations in the first and second halves of the 2006–2022 time frame. A total of 28235 citations were recorded from 2006 to 2022 and peaked at 5326 citations in 2017. As of May 10, 2022, word embedding and text augmentation publications were cited 170 times.

In 2006, two papers were published in the word embeddings research area. From then, the number of publications increased significantly, where the peak recorded in 2021 with 414 papers. In 2022, the published paper quantity is currently 110, but the year's total quantity is not yet final.

Figure 7 shows the average number of citations by each year in the period 2006–2022. It is calculated by dividing total publications by total citations. The peak result corresponds to publications made in 2013, where three articles collected 789 citations, which averages to approximately 263 citations per paper. After the climax in 2013, average citations sharply declined in the proceeding year. In 2017, we had 5326 citations, and the average citation was 28. As of May 2022, citation per paper is 1.55 (110 publications cited 170 times).

**Table 1.** Statistical characteristics of retrieved publications.

| Characteristics | Statistics |
|---|---|
| Total number of publications | 2050 |
| Number of unique publication sources | 869 |
| Number of unique countries/first countries | 88/83 |
| Number of unique affiliations/first affiliations | 2148/1237 |
| Number of unique authors/first authors/last authors | 4894/1629/1595 |
| Average number of citations | 11.95 |
| Average number of countries/regions in one publication | 1.19 |
| Average number of affiliations in one publication | 1.94 |
| Average number of authors in one publication | 3.51 |
| Average number of funds in one publication | 0.52 |
| Average number of pages in one publication | 17.94 |
| Average number of references in one publication | 33.46 |
| Average number of author keywords | 4.95 |
| Average number of words/characters in title | 10.34/81.57 |
| Average number of words/characters in abstract | 196.00/1,347.51 |

Additionally, we constructed a regression analysis based on the collected data. As independent variables, we picked $year/1000$ and $(year/1000)^2$, and the quantity of publications is assigned as the dependent variable. Our estimated regression model is: ($\hat{y} = 14903004 - 14833349 * (year/1000) + 3691000 * (year/1000)^2$). $R^2$ of the resulting model equals 0.958. In 2017 we had 415 publications, and the regression model predicted 397 publications.

Figure 8 shows Spearman correlation between four variables. It is observed that there exists moderate relation between Year and Total publication by year. As $R^2$ metric of the constructed regression model tells, the number of publications is explained by year by 95.8%. A positive correlation means that the number of publications rises each subsequent year.

### 3.2 Top Conferences and Journals

We analyzed total publications, total citations, and averages by conferences and journals and selected the top ten of them as the most contributing sources. Table 2 lists the top 10 conferences and journals that generated the most citations per publication (ACP). Two papers presented at the 52nd Annual Meeting of the Association for Computational Linguistics in 2014 were cited 905 times, resulting in 452.5 citations per paper for the conference. Article [16] titled "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification" from

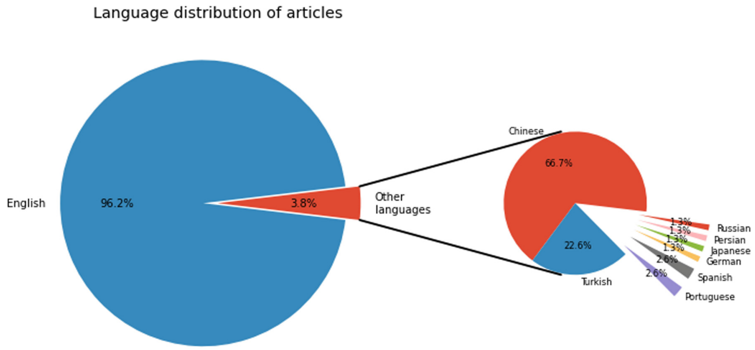

**Fig. 2.** Language distribution and most frequent words in the retrieved articles (in keywords, titles, and abstracts).

this conference was cited 837 times. Among journals, the Journal of Chemical Information and Modeling, with 177 citations, is worth noting. The article titled "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition" by Sabrina Jaegar et al. [8] was the top performing article in that journal.

### 3.3 Top Institutions and Funding Organizations

The Paper also analyzes institutions and funding organizations that help researchers publish papers in NLP related to word embedding and text augmentation. The results are illustrated in Fig. 9. In the top six, we observe only Chinese institutions and organizations; out of six, three are Chinese universities: Beijing University, Tsinghua University, and Peking University. The three universities have made a total of 88 publications in the period 2006–2022. Another interesting observation is that the list of top 10 institutions and organizations consists of only Asian universities and organizations, seven of which are from China, two from India, and one from Hong Kong SAR.

Next, Fig. 10 contains the top 10 foundations involved in investing in and sponsoring research paper publications. In the top 3 are two Chinese foundations: the National Natural Science Foundations of China and the National Key Research and Development Program of China. In total, they made contributions to the publication of 300 research papers. Besides Chinese organizations, in the top 10, we see two United States foundations: National Science Foundation and European Commission. In total, these two foundations have sponsored 109 paper publications.

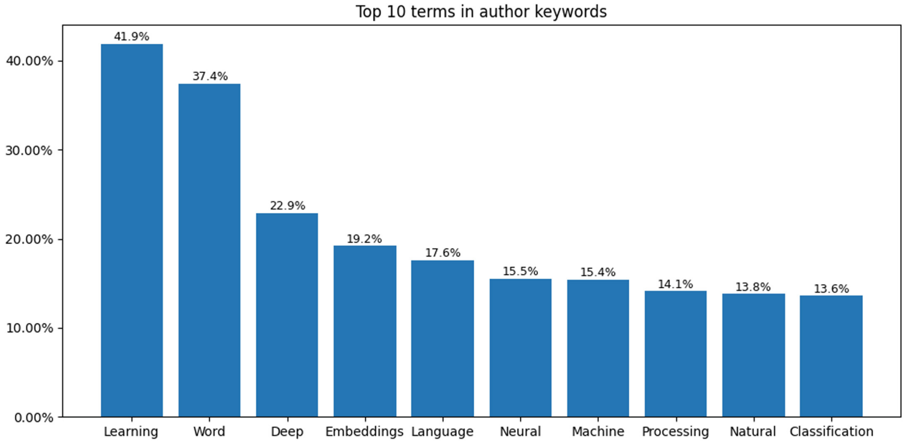

**Fig. 3.** Language distribution and most frequent words in the retrieved articles (in keywords, titles, and abstracts).

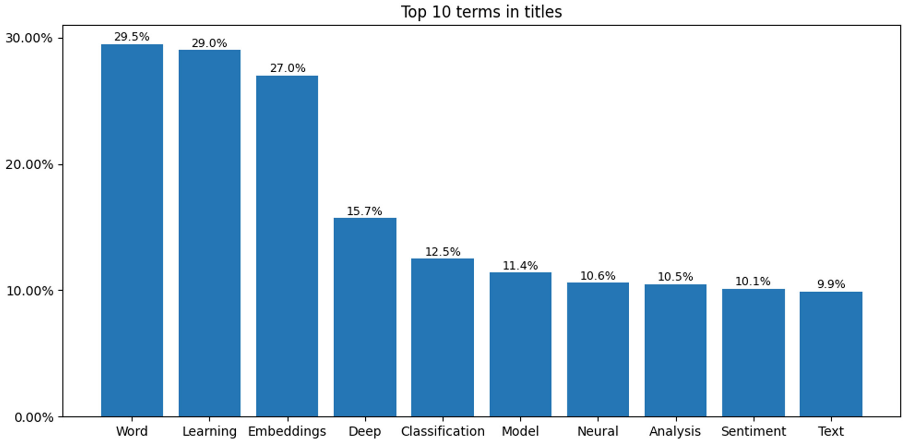

**Fig. 4.** Language distribution and most frequent words in the retrieved articles (in keywords, titles, and abstracts).

### 3.4 Top Influential Publications

Citations measure the success of the publications and their garnered interest. The more citations paper acquires, the more successful and influential it can be considered. Citation information from the dataset was analyzed, and the top 10 publications with the most citations were identified. Table 3 lists paper titles, authors' names, publication year, and total citations of the 10 most influential publications. A research paper titled 'Supervised learning of universal sentence

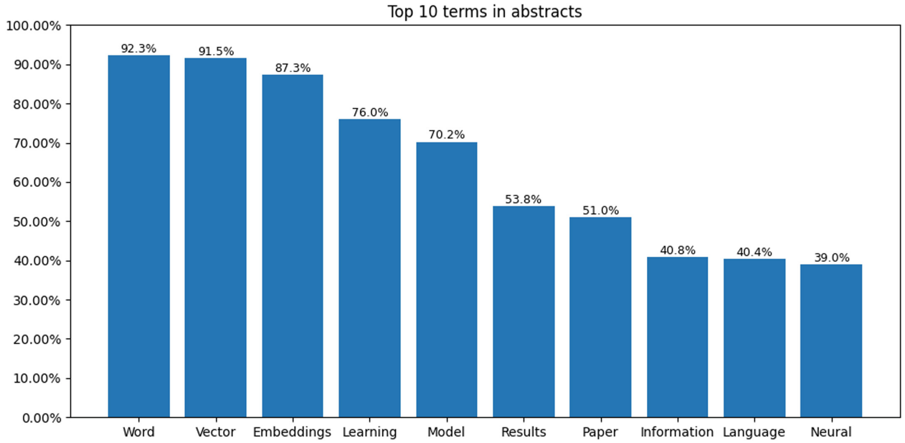

**Fig. 5.** Language distribution and most frequent words in the retrieved articles (in keywords, titles, and abstracts).

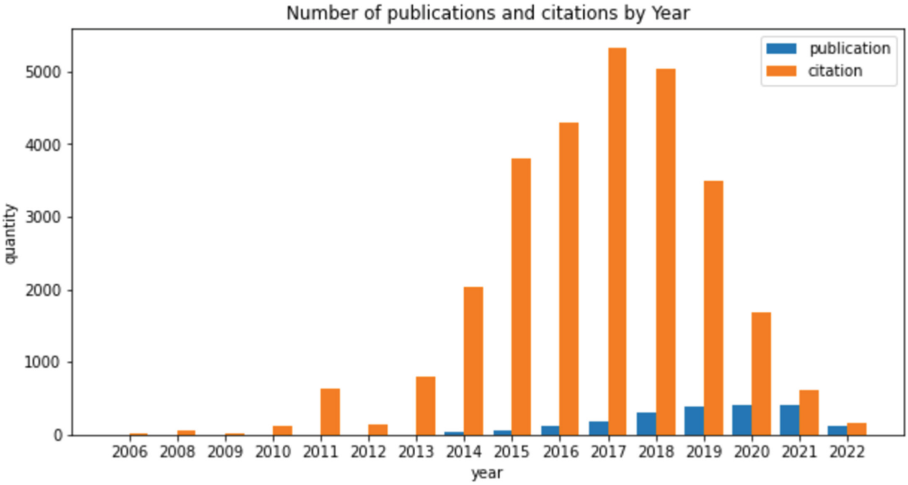

**Fig. 6.** Total publications and total citations distribution in 2006–2022 period.

representations from natural language inference data' by Conneu et al. [5], published in 2017, is the most influential work with a total of 909 citations. The top 10 is concluded with an article by Bordes, A. et al. and has 416 citations.

After further analysis, we found that there are 256 publications with more than 20 citations, 112 with more than 50 citations, and 56 papers with more than 100 citations from 2006 to 2022.

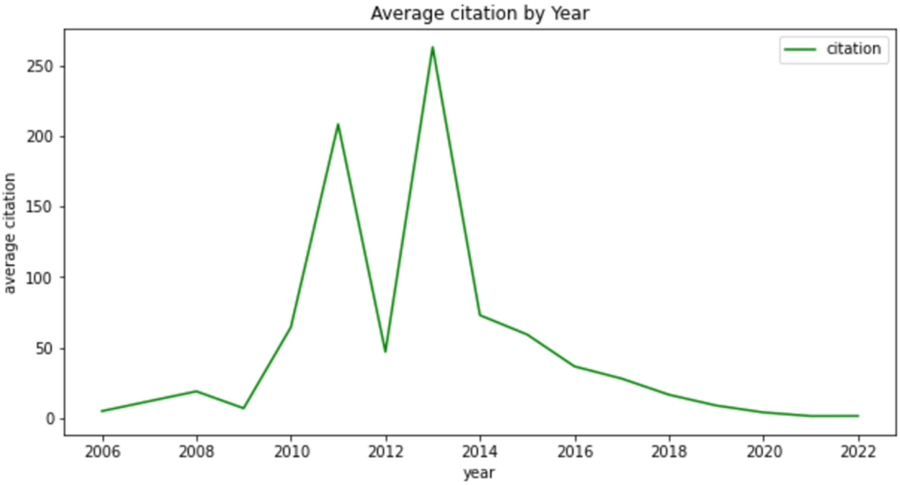

**Fig. 7.** Average citation per publication from 2006 to 2022.

**Table 2.** TP - total publication, TC- total citation, ACP - average citations per paper.

| Conference/Journal title | TP | TC | ACP |
|---|---|---|---|
| 52nd Annual Meeting of the Association for Computational Linguistics | 2 | 905.00 | 452.50 |
| Nature | 1 | 300.00 | 300.00 |
| 2nd International Conference on Learning Representations, | 1 | 260.00 | 260.00 |
| Advances in Neural Information Processing Systems | 7 | 1,546.00 | 220.86 |
| 10th ACM Conference on Recommender Systems | 1 | 189.00 | 189.00 |
| 20th SIGNLL Conference on Computational Natural Language Learning | 1 | 177.00 | 177.00 |
| Journal of Chemical Information and Modeling | 1 | 171.00 | 171.00 |
| 2014 Conference on Empirical Methods in Natural Language Processing | 2 | 327.00 | 163.50 |
| Synthesis Lectures on Human Language Technologies | 2 | 312.00 | 156.00 |
| 2017 IEEE International Conference on Software Quality, Reliability and Security | 1 | 155.00 | 155.00 |
| 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference | 4 | 601.00 | 150.25 |

### 3.5 Top Countries by Publications

An analysis of the number of publications by country was performed. In Fig. 11, bar chart depicts top-10 countries by paper publication quantity. 87 different

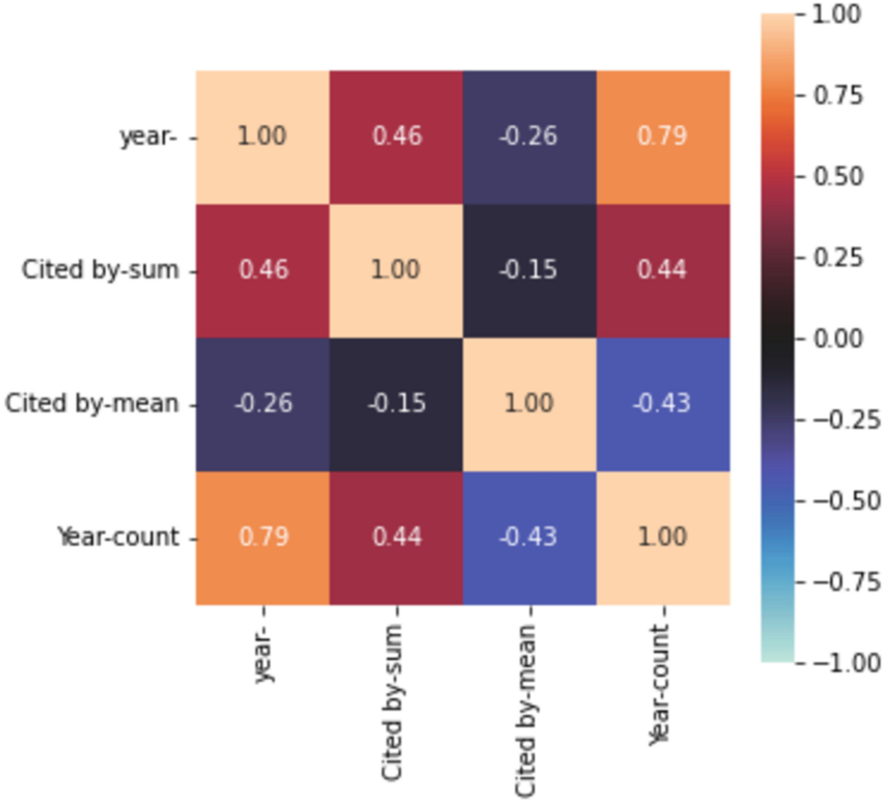

**Fig. 8.** Spearman correlation matrix of four variables. Correlation values range in [−1,1]. −1 and 1 correspond to the highest negative and positive correlations. In the figure, a high correlation between 'year' and 'Year-count' variables is observed. year - year of publications, Year-count is total publications by year, Cited by-mean - average citation value by year, Cited by-sum - total citation by year.

countries published two thousand fifty papers, and Kazakhstan has two publications among these. It is observed that China is in the leading position with more than 500 publications. This result is consistent with other analyses presented earlier, i.e., in influential publications, institutions, and funding organizations ranking, China is also holding the top position. The USA's next position is owned by the list, contributing 350 publications from 2006 to 2022. And India comes in 3rd place with 215 papers.

In the most influential publication list, papers published by China appear four times in the top 7. With a total of 2423 citations, it results in an average of 606 citations per paper.

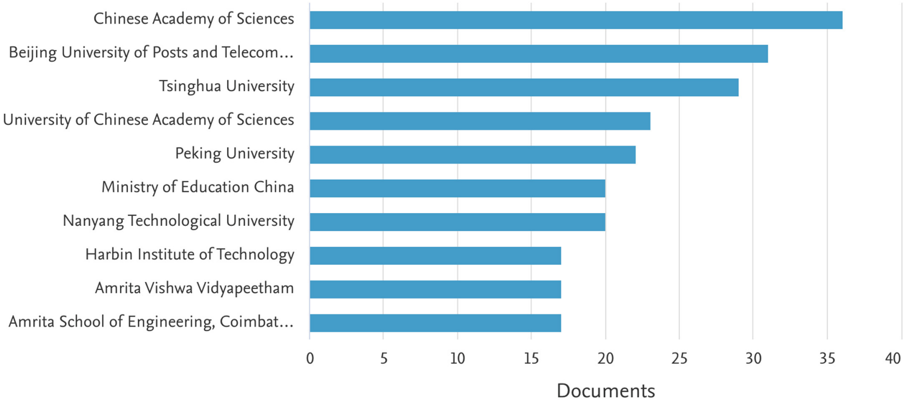

**Fig. 9.** Institutions and organizations listed in top 10 on paper publications.

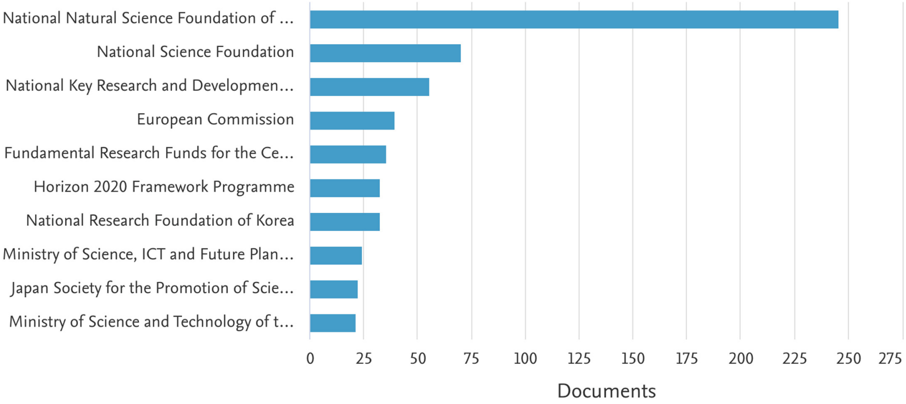

**Fig. 10.** Top 10 foundations sponsoring paper publications. The horizontal axis represents the number of publications sponsored by the foundation.

## 4 Discussion

The study provides the most recent bibliometric analysis on the topic of corpora generation for learning word vector embeddings. The research is based on the publication data from 2006 to 2022 retrieved from Scopus scientific database. Key findings are included in this section.

There is an upward trend in the number of publications made in the research area of the paper. In 2006, which corresponds to the lowest value of the metric, 2 papers were published. Since then, it has been on a steady rise, and 2021 corresponds to the peak number of publications which equals 414.

However, the average number of citations does not portray a significant upward trend from 2006 to 2022 because both the number of citations and the number of publications are increasing simultaneously, although at different rates.

**Table 3.** Top 10 influential publications based on total citations.

| Title | Author | Year | Citations |
|---|---|---|---|
| Supervised learning of universal sentence representations from natural language inference data | Conneau, A. et al. | 2017 | 909 |
| Learning sentiment-specific word embeddings for Twitter sentiment classification | Tang, D. et al. | 2014 | 837 |
| Zero-shot learning through cross-modal transfer | Socher, R. et.al | 2013 | 743 |
| Graph embedding techniques, applications, and performance: A survey | Goyal, P. et al. | 2018 | 698 |
| Is Man to computer programmer as woman is to homemaker? Debiasing word embeddings | Bolukbasi, T. et.al | 2016 | 658 |
| PTE: Predictive text embeddings through large-scale heterogeneous text networks | Tang, J. et.al | 2015 | 475 |
| Deep Sentence embeddings using long short-term memory networks: Analysis and application to information retrieval | Palangi, H. et.al | 2016 | 453 |
| A simple but though-to-bit baseline for sentence embeddings | Arora, S. et.al | 2017 | 437 |
| Learning deep representations of fine-grained visual descriptions | Reed, S. et.al | 2016 | 417 |
| Learning structured embeddings of knowledge bases | Bordes, A. et.al | 2011 | 416 |

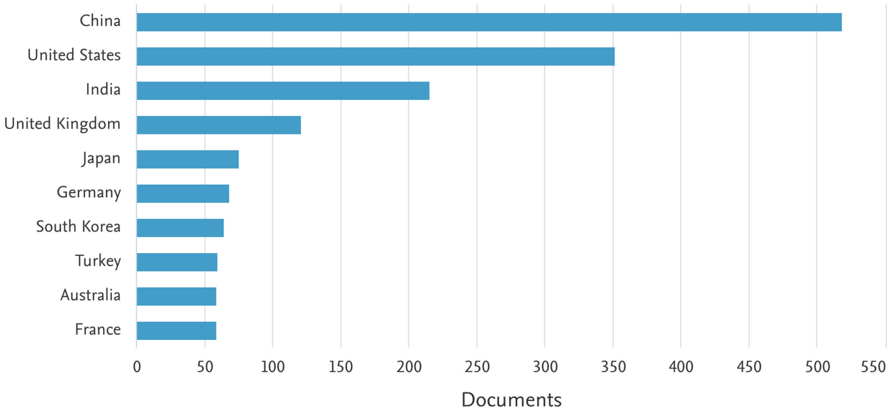

**Fig. 11.** Top 10 countries by publications. The horizontal axis represents the number of publications for each country from 2006 to 2022.

Instead, its graph demonstrates fluctuations, and a peak value of 263 citations in 2013 is easily noticeable.

Key statistical characteristics show that collected 2050 publications come from 869 unique sources, and 4894 unique authors with 2148 unique affiliations have contributed to the publications.

Top 3 most influential publications sources (conferences and journals) include *52nd Annual Meeting of the Association for Computational Linguistics*(ACP 452.5), *Nature* (ACP 300), and *2nd International Conference on Learning Representations* (ACP 260). Such high values of average citations are explained by the small number of articles (range in 1–7) from the sources and high citation numbers of the articles in the top-10 publication sources.

The most influential publication in terms of citations was ‘Supervised learning of universal sentence representations from natural language inference data’ by Conneu et al. [5], and it was cited 909 times. The number of citations in the top 15 most influential papers (published between 2011 and 2019) ranges between 260 and 909. Ten of the articles from the list were written in collaboration between 4 and more authors (up to 9 authors). Only one article titled”Neural Network Methods for Natural Language Processing” was written by a single author [7].

Another interesting observation was made by comparing the results of analyses conducted to identify the top institutions and funding organizations and top countries by publications, presented in Sects. 3.3 and 3.5, respectively. The top 6 institutions and organizations from the list of most influential institutions are from China. Also, 2 foundations from China are among the top 10 paper publications sponsoring foundations, placed first and third. And these results are consistent with the ranking of 87 countries by publications. China leads the list with more than 500 (out of 2050) publications made from 2006–2022. In addition, 4 papers from the list of 10 most influential publications are from China. These results show the expertise of China in the research area this paper covers.

## 5 Conclusion

The research conducted is bibliometric analysis in NLP area, specifically on learning vector analysis and text augmentation methods. Data generation on bibliometric analysis was acquired from Scopus, where the dataset consists of papers published from the period 2006–2022. We made a descriptive analysis on the 2050 paper to gain knowledge on statistical and general descriptive information. By analyzing, we get the results on total publication, total citation, and average citation per publication distributed in a timeframe. We found China as the major contributor to the top list of most influential journals, organizations, publications, and institutions.

In the end, we believe that the compiled bibliometric research would help researchers to get knowledge on general descriptive features, patterns, and insights to systematically design their studies and get the best results on research.

**Acknowledgments.** This research is conducted within the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan under the grant number AP09260670 “Development of methods and algorithms for augmentation of input data for modifying vector embeddings of words.”

## References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. CoRR arXiv:abs/1607.04606 (2016)
2. Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: a computational study. Behav. Res. Methods. **39**, 510–526 (2007)
3. Chakma, K., Das, A.: CMIR: a corpus for evaluation of code mixed information retrieval of hindi-english tweets. Computacion y Sistemas. **20**, 425–434 (2016). https://doi.org/10.13053/CyS-20-3-2459
4. Chiu, W.T., Ho, Y.S.: Bibliometric analysis of tsunami research. Scientometrics **73**, 3–17 (2007). https://doi.org/10.1007/s11192-005-1523-1
5. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. CoRR arXiv:abs/1705.02364 (2017)
6. Geng, Y., et al.: A bibliometric review: Energy consumption and greenhouse gas emissions in the residential sector. J. Clean. Prod. **159**, 301–316 (2017). https://doi.org/10.1016/j.jclepro.2017.05.091
7. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers, San Rafael (2017)
8. Jaeger, S., Fulle, S., Turk, S.: Mol2vec: unsupervised machine learning approach with chemical intuition. J. Chem. Inf. Model. **58**(1), 27–35 (2018). https://doi.org/10.1021/acs.jcim.7b00616. pMID: 29268609
9. Merigo, J.M., Gil-Lafuente, A., Yager, R.: An overview of fuzzy research with bibliometric indicators. Appl. Soft Comput. **27**, 420–433 (2015). https://doi.org/10.1016/j.asoc.2014.10.035
10. Merigó, J.M., Gil-Lafuente, A.M., Yager, R.R.: An overview of fuzzy research with bibliometric indicators. Appl. Soft Comput. **27**(C), 420–433 (2015). https://doi.org/10.1016/j.asoc.2014.10.035
11. Neuhaus, C., Daniel, H.D.: Data sources for performing citation analysis: an overview. J. Document. **64**, 193–210 (2008). https://doi.org/10.1108/00220410810858010
12. Ojo, O.E., Ta, T.H., Gelbukh, A., Calvo, H., Sidorov, G., Adebanji, O.O.: Automatic hate speech detection using deep neural networks and word embedding. Computacion y Sistemas. **26**(2), 1007–1013 (2022). https://doi.org/10.13053/CyS-26-2-4107
13. Sasaki, S., Suzuki, J., Inui, K.: Subword-based compact reconstruction for open-vocabulary neural word embeddings. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3551–3564 (2021). https://doi.org/10.1109/TASLP.2021.3125133
14. Shekhar, S., Sharma, D., Beg, M.: An effective BI-LSTM word embedding system for analysis and identification of language in code-mixed social media text in English and roman Hindi. Computación y Sistemas. **24**, 1415–1427 (2020). https://doi.org/10.13053/cys-24-4-3151

15. Singla, K., Bose, J., Varshney, N.: Word embeddings for IoT based on device activity footprints. Computación y Sistemas. **23**, 1043–1053 (2019). https://doi.org/10.13053/cys-23-3-3276
16. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for Twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1555–1565. Association for Computational Linguistics, Baltimore, Maryland, June 2014. https://doi.org/10.3115/v1/P14-1146
17. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. CoRR arXiv:abs/1509.01626 (2015)