Research Trends and Applications of Data Augmentation Algorithms

Joao Fonseca[1*], Fernando Bacao[1]

[1]NOVA Information Management School, Universidade Nova de Lisboa

*Corresponding Author

Postal Address: NOVA Information Management School, Campus de Campolide, 1070–312 Lisboa, Portugal

Telephone: +351 21 382 8610

1 Introduction

Introduction goes here.

2 Theory

This part is going to be the very last thing to be done, since it doesn't seem absolutely necessary.

3 Methodology

In this section we describe the procedures defined for the literature collection, data preprocessing and literature analysis. The analysis of the literature was developed with 3 different approaches. Throughout the analyses, data preprocessing and hyperparameter tuning was developed iteratively. The procedure adopted in this manuscript is shown in Figure 1.

The analysis and modelling was developed using the Python programming language, along with the Scikit-Learn [1], Gensim [2], Umap-Learn [3] and Networkx [4] libraries. The final network analysis and visualization was done with Gephi [5].

An exploratory data analysis was done to understand which manuscripts, journals and conferences are most significant within the field of Data Augmentation.

Text Mining perspective

- General overview of the steps taken to produce the visualizations and analyses.

- Use diagrams and visualizations.

- Discuss embedding techniques used.

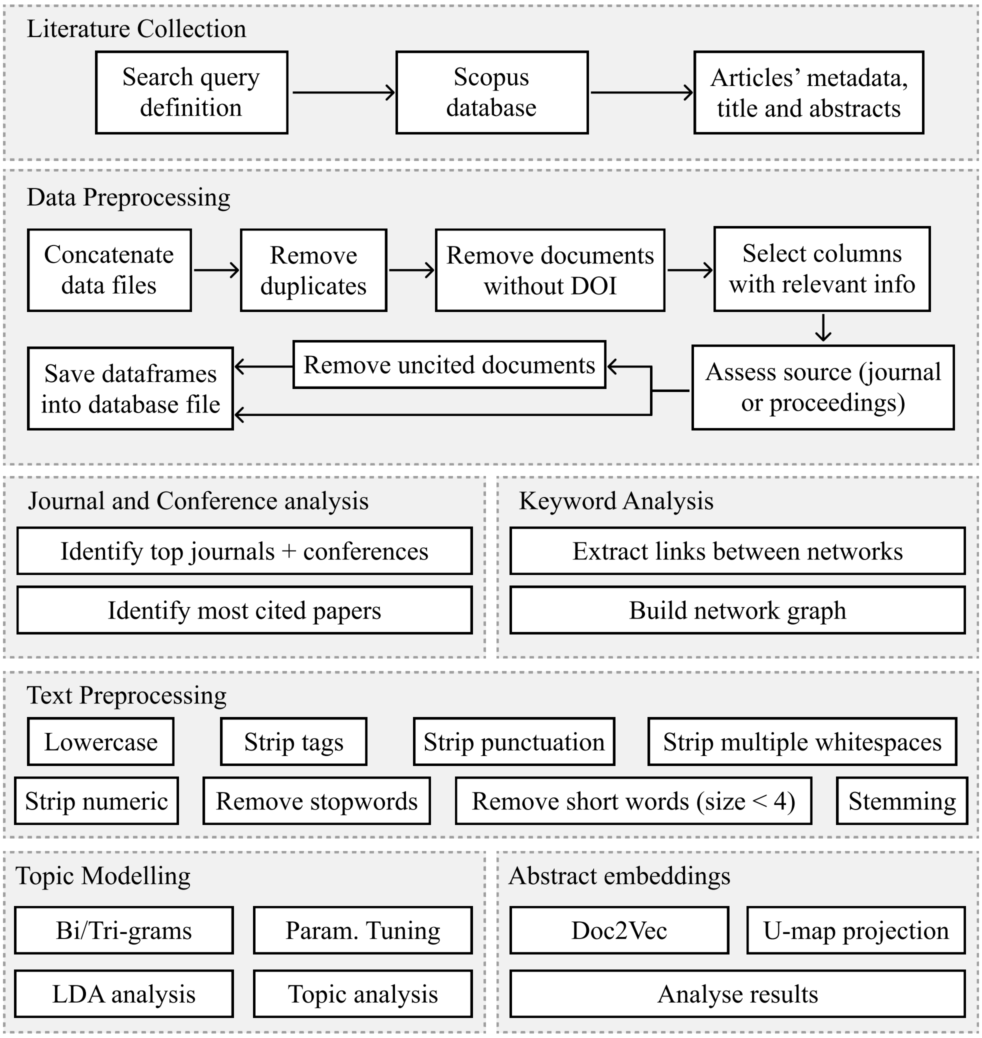

Figure 1: Diagram of the proposed literature analysis approach.

3.1 Keyword Identification and Search

The focus of this literature analysis is to understand the different algorithms, domains and/or tasks that employ data augmentation techniques. Therefore, we use the keyword "data augmentation" in order to ensure an unbiased analysis. The results were then limited to conference papers and journal articles written in English that were published in the past 15 years. Due to the already large amount of results found, the document retrieval was done using solely the Scopus database. The resulting query is shown below:

```
KEY ( "data augmentation" )  AND  ( LIMIT-TO ( LANGUAGE ,  "English" ) )
AND ( LIMIT-TO ( DOCTYPE ,  "cp" )  OR  LIMIT-TO ( DOCTYPE ,  "ar" ) )
AND  (
        LIMIT-TO ( PUBYEAR ,  2021 )  OR  LIMIT-TO ( PUBYEAR ,  2020 )
    OR  LIMIT-TO ( PUBYEAR ,  2019 )  OR  LIMIT-TO ( PUBYEAR ,  2018 )
    OR  LIMIT-TO ( PUBYEAR ,  2017 )  OR  LIMIT-TO ( PUBYEAR ,  2016 )
    OR  LIMIT-TO ( PUBYEAR ,  2015 )  OR  LIMIT-TO ( PUBYEAR ,  2014 )
    OR  LIMIT-TO ( PUBYEAR ,  2013 )  OR  LIMIT-TO ( PUBYEAR ,  2012 )
    OR  LIMIT-TO ( PUBYEAR ,  2011 )  OR  LIMIT-TO ( PUBYEAR ,  2010 )
    OR  LIMIT-TO ( PUBYEAR ,  2009 )  OR  LIMIT-TO ( PUBYEAR ,  2008 )
    OR  LIMIT-TO ( PUBYEAR ,  2007 )  OR  LIMIT-TO ( PUBYEAR ,  2006 )
)
```

Started with a single keyword: "Data Augmentation", 4618 results.

Limited results to documents written in English, 4517 results.

Only include articles and conference papers, 4443 results.

Consider papers published in the past 15 years, 4281 results.

Final query:

Due to the limitations in the Scopus data export (maximum 2000 documents per export), the data was extracted in four parts: 2021, 2020, 2019 and 2018 — 2006.

One of the exported references had a corrupted line, which caused the loss of one additional document.

Removed references without a DOI, 3948 results.

Removed references without a single citation, 2259 results. Exception made for journals and conference analysis.

For the network analysis:

Out of the 2259 results, 1923 contained keywords in Scopus' database.

Keyword combinations showing up in only one document are removed from further analysis.

The network consists of an undirected graph whose weights consist of the following formula: $Keyword--pairweight = \log(Avgcites * Nbrofdocuments)$ to avoid a disproportional bias of highly cited research articles.

3.2 Repositories

- Repositories: Scopus

3.3 Bibliometric Analysis

- Discuss data preprocessing done.
- Concatenate data, extract features (if possible), data cleaning etc

3.4 Data Analysis Software for Bibliometric Research

- Python
- Streamlit
- Text Mining and Embedding libraries
- Data Visualization libraries
- VOSviewer?

4 Results

Results go here.

4.1 PRISMA Flow Diagram

- Create a flowchart with the data cleaning process and describe it.

4.2 Terms Frequency

4.3 Topics Discovered

LDA analysis goes here

4.3.1 Main Journals

4.3.2 Main Conference Proceedings

4.4 Author Co-occurrence Analysis

Not sure whether to keep this one.

4.5 Title and Abstract Text Occurrence Analysis

This can be done by text occurrence network visualization or embedings

4.6 Most Cited Publications

4.7 Application and Method Analysis

5 Discussion

Discussion goes here.

5.1 Research Question Discussion

5.2 Research Gap Discussion

5.3 Study Limitation Discussion

6 Conclusions

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[2] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010. http://is.muni.cz/publication/884893/en.

[3] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[4] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring Network Structure, Dynamics, and Function using NetworkX," in *Proceedings of the 7th Python in Science Conference* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11–15, 2008.

[5] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 3, 2009.