Text Analysis of Scientific Publications

Amanda Agyeiwaah
University of Passau
agyeiw01@ads.uni-passau.de

Bhaumi Panchal
University of Passau
pancha01@ads.uni-passau.de

Umme Rabab
University of Passau
rabab01@ads.uni-passau.de

KEYWORDS

Text analysis, Topic modelling, Unsupervised algorithm : LDA(Latent Dirichlet Allocation), Text classification, Supervised algorithm : LSTM(Long Short-Term Memory)

1 ABSTRACT

The world wide web provides an endless source of knowledge, in which an Information extraction and finding relevant links between the scientific publication is a very complex task in the exponential growth of the information space. The volume of the scientific literature increases exponentially and it creates critical problems for researchers and students who want to quickly and easily find the literature of specific topics among a large number of scientific publications. To improve this problem, this paper proposes a Text classification approach for supervised algorithms and a Topic modelling approach for an unsupervised algorithm which are based on Natural Language Processing Techniques. We used the LSTM (Long Short-Term Memory) model for text classification and LDA (Latent Dirichlet Allocation) model for Topic modelling. Every document is classified into a topic in the text classification, and for the topic modelling,it identify the latent topics in the documents. We use the *ACL Anthology Reference Corpus*[1] in our experiment for text classification and topic modelling.

2 INTRODUCTION

It is an undeniable fact that the internet is plagued with many scientific papers, journals, conferences and articles. This is because day in and day out there are new research areas into interesting topics ranging from medicine, IT, agriculture, literature, entertainment, etc. The thirst for knowledge and advancement of technologies call for research and publications of scientific findings. This accounts for the ever-growing number of scientific papers published in various fields. These days students and researchers find themselves overwhelmed by the huge and rapid growth of scientific publications on the world wide web especially during literature analysis and assessment to find relations between studies in diverse domains.

Anytime there is the need to do research, one cannot shy from sampling a few research papers that have a bearing or aligns with the topic under study as this allows for easy comparison and drawing of motivation, approaches, methodologies and conclusions from various research literature to make one's own assessment; something we call information gathering in a particular field. This knowledge gathering becomes difficult when it is done entirely manually. Extracting information and establishing their syntactic and semantic meaning in these scientific publications can be daunting hence the need to apply NLP techniques to achieve semi-automatic knowledge extraction and classification methods relevant for decreasing time and effort spent on extracting information from scientific papers thus increasing efficiency. With the help of NLP techniques, only essential concepts are captured. When this is achieved, it will go a long way to help students and researchers sample any kind of research papers they want and easily extract the right information for the right purposes.

3 PROBLEM STATEMENT

The influx of scientific publications have brought the need to distill information from scientific papers and to deduce meaning of the texts that make up these papers. During research, data collected must be analysed to find if any relationship exists between data or if it can be used to make an inference to get a better understanding.

How is text examined to identify themes and trends that can enable the researchers take strategic action? The use of NLP and Machine learning techniques to find meaning in big texts eradicates the issue with time-consuming, inefficiency and inaccuracy.

3.1 Objectives

To analyse and categorize texts in scientific papers.

4 INITIAL PLAN

Claims ought to be argumentative. Hence, a text or phrase qualifies to be a claim if you have to argue for a certain clarification or understanding of your subject matter. A good claim has to be specific. We are more interested in scientific claims which follow the natural language sentence structure by demonstrating the relationship that exist between two entities [10]Thus, how does one affect the other? In this paper, we propose an extension of [7] by:

- Identifying and extracting claims from scientific papers.
- Finding rhetorical relations that exist
- Performing clustering analysis of scientific claims to find similar and dissimilar relations
- Explore if an expression makes meaning by depending on another text called antecedent(presence of anaphora-a linguistic evidence). Example of Anaphora: Hanson dropped the ball. It bounced back quickly - The pronoun 'it' is an anaphor; it points to the left toward its antecedent the ball.

4.1 Changes in Initial Plan

The above initial plan started on a good course however finding how the RST technique was really implemented was a challenge given the time constraints. As a result of this challenge, there was the need to make some changes and modifications to the original plan whiles still sticking to scientific publications. Hence we switched to Text Analysis of Scientific Publications which also fits into the original scope.

[1] https://web.eecs.umich.edu/~lahiri/acl_arc.html

Univeristät Passau, Text Mining Project, 2020
.

5 METHODOLOGIES

NLP based techniques are used for the automatic extraction of meaningful data from texts and get sensible information from word to paragraphs to documents. Text analysis is an NLP based technique that is used to handle a large amount of data and assigning tags or classification to each scientific literature. There are two approaches for the labels and classification of datasets :

(1) **Topic modelling :** Topic modelling is one such technique, at the document level, which can extract meaningful information in the form of latent semantic topical structures from a collection of documents[6]. Contrasting from rule-based text mining techniques, it is an unsupervised technique used for finding a collection of words, i.e., a topic, such that the words exist in a repeating pattern of co-occurrence in a corpus of documents[6].

Various methods are used to obtain topic modelling such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM). This paper represents the Latent Dirichlet Allocation (LDA) model for topic modelling since it is highly studied and the most popular model in many domains and numerous toolkits such as Machine Learning for Language Toolkit (MALLET), Gensim, and Stanford TM toolbox (TMT) because it can address other models' limitations, such as latent semantic indexing (LSI) and probabilistic latent semantic indexing (PLSI)[2].

Latent Dirichlet Allocation (LDA) : Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar[11]. Each document consists of different topics, where each topic is a probability distribution over words[2]. We will know the involvement of each topic from the probability distribution. The main benefit of using the LDA model is that it does not require any input from previous knowledge, topics can be directly assumed from a given dataset. For example, the 'Animals' topic has word “dog”, ”cat”,” chicken”,” zoo” with the high probability and the cooking topic has the word “oven”,” food”,” taste”,” restaurants”,” plates” with the high probability. Then, each topic has a probability distribution over words.

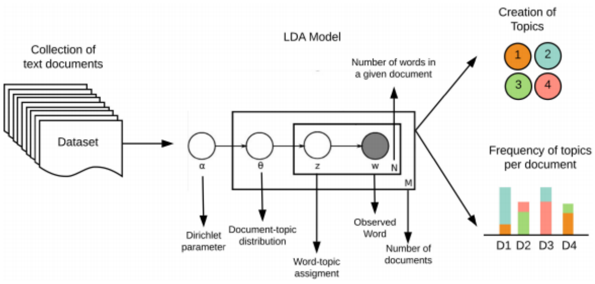


Figure 1: Schematic diagram of LDA algorithm.[5]

(2) **Topic Classification :** Topic classification is a supervised machine learning task, in which topics are predefined. After a list of topics and reliable datasets, we can train the topic classification algorithm by tagging the text data with the predefined list of topics.

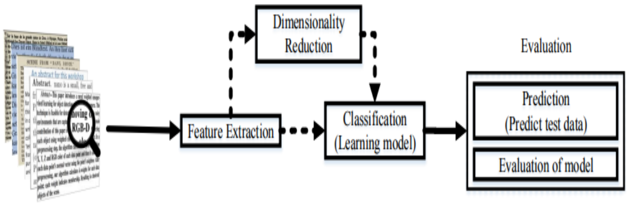


Figure 2: Text Classification Pipeline.[8]

Text classification-based machine learning technique includes four steps:

(a) Data Pre-processing : Convert row data into a clean dataset.
(b) Feature Extraction : After data has been cleaned, Feature Extraction method can be applied to extract the features from raw data.
(c) Model Training : Machine learning model is trained on a labelled dataset.
(d) Model Evaluation : Evaluate model performance.

In this paper, different supervised machine learning models are used for text classification such as Long Short-Term Memory(LSTM), Naive Bayes and SVM (Support vector machine).

Long Short-Term Memory(LSTM):

LSTM is a deep learning algorithm that adds context into word vectors via Recurrent Neural Networks. The choice of Recurrent Neural Networks include:

- The fact that it uses sequential information.
- It keeps in memory previous calculations and uses for prediction of next events.
- It is ideal for text analysing.

Multinomial Naive Bayes : Multinomial Naive Bayes(MNB) classifier is a type of NB classifier and is very useful in natural language processing for prediction and classification purposes.

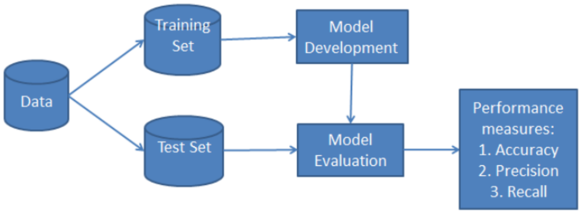


Figure 3: MNB Model[1]

There are two classification stages such as the learning phase and the Evaluation phase.At the learning stage, classifier trains its model on a given dataset and in the evaluation phase, it tests the classifier performance[1].Performance is evaluated on the based on various parameters such as accuracy, error, precision, and recall rate[1].then, data is spit into training and testing dataset.

Support Vector Machine(SVM) : SVM is widely used in text classification problem. Data pre-processing and feature engineering steps are the same as other algorithms. The processing steps of SVM algorithm are the same as multinomial Naïve Bayes, However, unlike calculating the frequency of each feature in the document in Multinomial Bayes, SVM plots each feature vector as an n-dimensional point (where n is the number of total features in a post) with the value of each feature corresponding to the value of a particular coordinate[12]. Then, we perform classification by finding the hyperplane that separates points from the different classes ("support vectors" are the points near the boundary hyperplane)[12]. Thus, the core problem of learning SVM classifier is to find the best separation hyperplane[12].

6 IMPLEMENTATION AND RESULTS

6.1 Text Classification

Many of the papers in the ACL corpus did not have abstracts hence the first 250 words were taken into consideration as the abstracts and used for classification. To maintain data format, data was converted to lower letters and stopwords removed along with punctuation marks. Top words were extracted by setting minimum occurrence to 5 so that if any word is present more than 5 times during execution, only this data is kept and saved to a vocabulary file which is processed again later. The following are the 8 labels done on basis of abstract as it defines the whole paper :

1. Speech: Papers categorized as speech
2. Translation: Papers categorized as translation theory
3. Information Extraction(IE): Papers categorized as information extraction or retrieval
4. Grammar: Papers categorized as Grammar
5. Lexical: Papers categorized as Lexical
6. QuestionAnswer(QA):Papers categorized as QA
7. Summarization: Papers categorized as Summarization.
8. Named Entity(NE): Papers categorized as Named Entity

The data was split into 70% training and 30% with shuffling set to true to avoid repetition of data that is being sent to the model. The models used were Long Short Term, Support Vector, and Multinomial Naïve Bayes. These models were selected because of their advantages with regard to text classification.

6.1.1 LSTM :. The following details the result obtained after a 100 epoch was reached during the training and testing using TensorFlow API and Keras. The models were trained on 310 training data and validated with 133 test data. The graph below shows the performance of the data with regard to LSTM. Here the model performed better than the others with an accuracy of 0.8650 which promises a good result. However during training, an issue of overfitting was encountered hence more data was included and regularization of certain hyperparameters such as learning rate, optimizers, and loss functions. The result showed that to improve the accuracy and make it more stable, more LSTM units must be added.

6.1.2 Multinomial Naive Bayes: Accuracy here is worst with 0.15. Neither is it classifying good.

6.1.3 SVM:. Much was expected from SVM as it is one of the best in terms of classifiers, but it gave poor results too.It was not able to make decision boundaries.

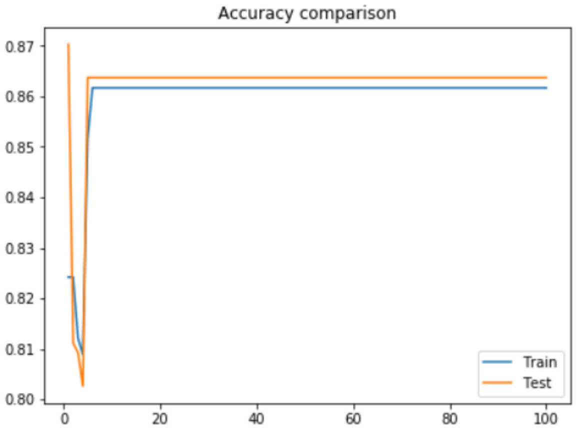


Figure 4: Accuracy Comparison.

LOSS COMPARISON : The loss function is the difference between true value and predicted value.

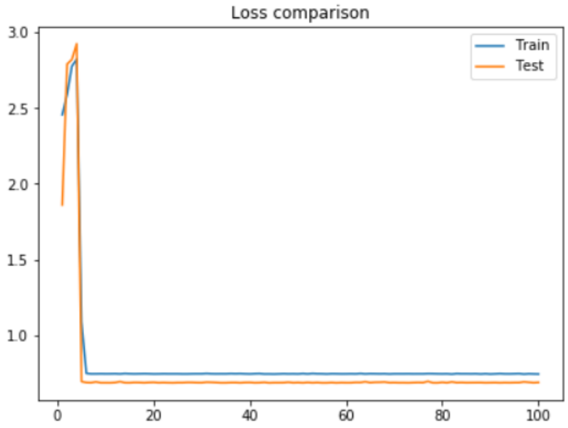


Figure 5: Loss Comparison.

CONFUSION MATRIX : The result of a bad confusion matrix arises because of model not classifying to expectations. For example, it can be seen that the SVM model misclassifies since it shows that 17 instances which actually belonged to grammar papers were misclassified and predicted as speech.

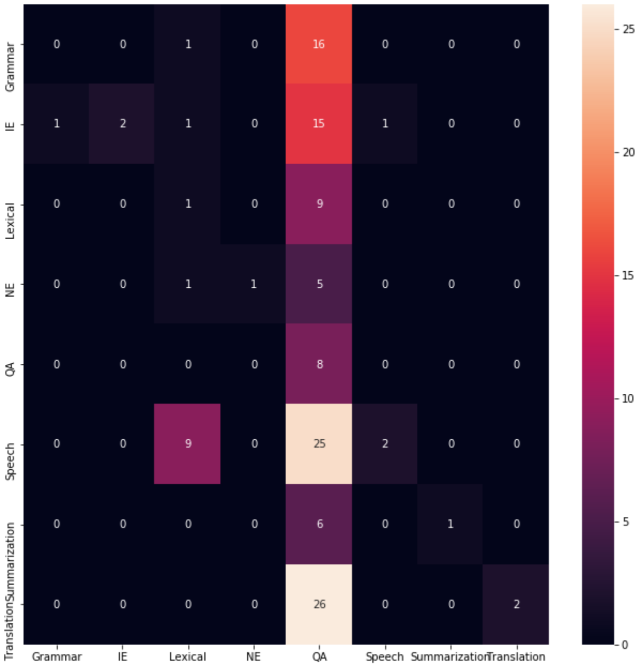


Figure 6: Multinomial Naive Bayes.

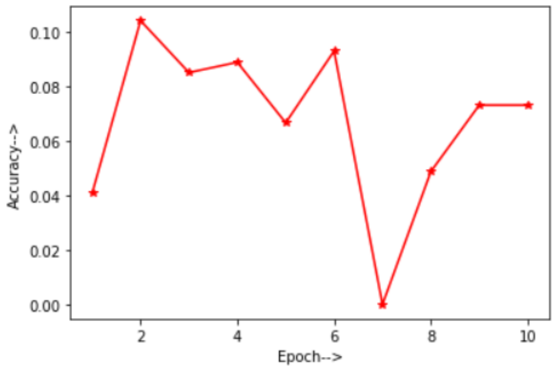
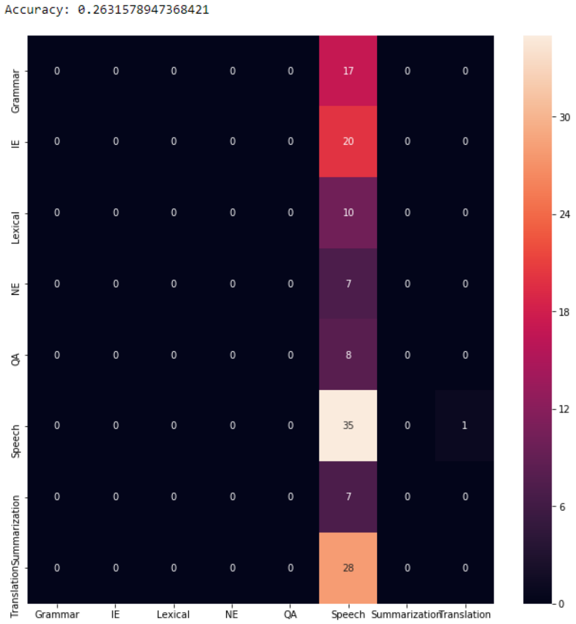


Figure 7: Naive Bayes Line Graph.



```
Accuracy: 0.2631578947368421
```

Figure 8: SVM.

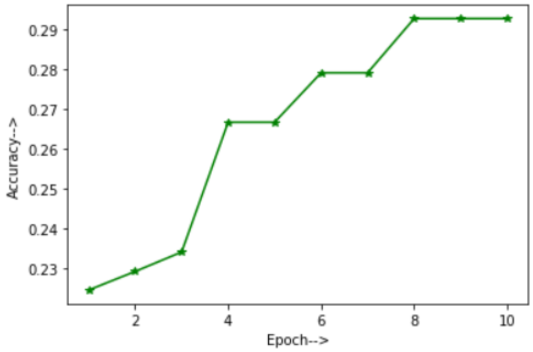


Figure 9: SVM Line Graph.

CROSS VALIDATION The graphs of Support Vector Machine and Naïve Bayes were plotted using cross validation as these do not store historical data. Cross validation will split data into 10 batches by using 9 batch for training and 1 batch for testing so here only 10 accuracy points can be plotted as shown in the figures displayed.

6.2 Topic Modelling

Topic modelling is a text analysis technique which helps finding hidden topics from a large corpus. It is implemented in three steps namely data cleaning, exploratory data analysis (EDA) on the cleaned documents and lastly extracting topics using Latent Dirichlet Allocation (LDA).

First 100 documents are chosen from the main corpus i.e. ACL Anthology Reference Corpus for our purpose in order to not burden the computation processes. Cleaning the data involves removing punctuation, numbers and lowering all characters in the document. The documents are converted into matrices in order to apply computation and EDA, namely a document term matrix (DTM) where rows constitute documents and columns are made up of all words in the corpus. A DTM represents the occurence of all words in each document. Documents can be visualized using a framework ’WordCloud’[9]. Below are generated word clouds from the first four documents shown in Fig.10 :

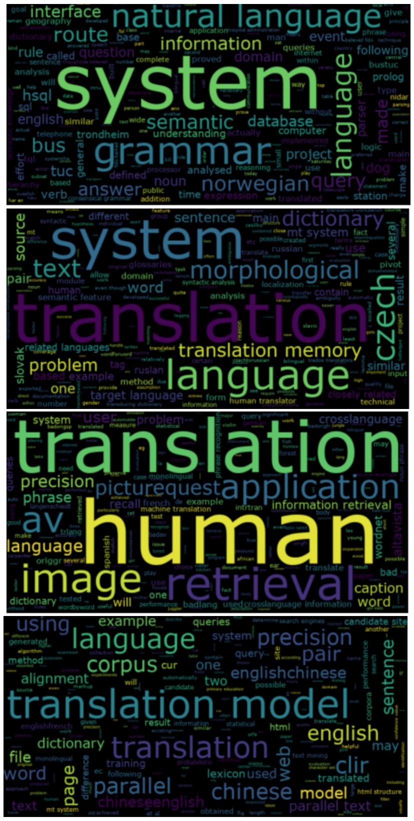


Figure 10: Word clouds generated from 4 documents

As can be seen, few words have more weight i.e. are more used in the document which gives us an idea of the topic category that the document migt belong.

6.2.1 Latent Dirichlet Allocation (LDA) :. Regarding topic modelling, LDA is the most popular probabilistic model which extracts topics from a collection of documents or corpus. Each document in the corpus is represented as a collection of topics and each topic is represented as a probability distribution over different words [3]. α and β are two hyper parameters of the LDA model, where α represents document-topic density and β represents topic-word density[4]. Other parameters required by the model are the number of topics and passes. Number of topics are set on the basis of how many topics one thinks are there in the corpus and the passes help the model learn through each iteration and can be tweaked for as long as the satisfactory topic distribution is not achieved.

We initially set the number of topics to 5 as well as the number of passes. Following is the topic distribution we achieved on initial run for 5 topics with highest coherence amongst the rest:
Topic 1 - 0.007*"information", Topic 2 - 0.005*"data", Topic 3 - 0.009*"grammar", Topic 4 - 0.006*"words", Topic 5 - 0.007*"word"
As can be seen, the results give us an idea of what the documents might be but are not satisfactory. A lot of factors lead to better results such as elimination of unwanted words from the corpus and tweaking the parameters of the LDA model.

7 CONCLUSION

In this paper, we have presented text analysis of scientific publications through in two ways namely text classification and topic modelling. We implemented text classification based supervised algorithms such as LSTM, Multinomial Naïve Bayes, SVM. Also implemented LDA for topic modelling which is an unsupervised algorithm. We achieved significant results in text classification performance with the help of LSTM model. LSTM model gave us the highest accuracy in comparison to other models.
For future work more attention to data cleaning and organizing can be given as that is the first step which then leads to better results later. Regarding LDA, increasing the number of topics and passes can stabilize the topic distribution over documents. The distribution can then be visualized and further determined which document belongs to which topic. Regarding text classification, future work will involve adding more labelled data to help boost the result which could help other models like SVM set better decision boundaries as well.

8 RESPONSIBILITIES

Table 1: Assignment of Responsibilities

Section	Responsible
1	Amanda Agyeiwaah, Bhaumi Panchal
2	Amanda Agyeiwaah, Bhaumi Panchal
3	Amanda Agyeiwaah, Bhaumi Panchal
4	Amanda Agyeiwaah, Bhaumi Panchal
5	Bhaumi Panchal
6	Bhaumi Panchal
7.1	Amanda Agyeiwaah
7.2	Umme Rabab
8	Amanda Agyeiwaah, Bhaumi Panchal, Umme Rabab

REFERENCES

[1] Muhammad Abbas, Kamran Ali Memon, Abdul Aleem Jamali, Saleemullah Memon, and Anees Ahmed. 2019. Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS* 19, 3 (2019), 62. https://www.researchgate.net/profile/Anees_Ahmed6/publication/334451164_Multinomial_Naive_Bayes_Classification_Model_for_Sentiment_Analysis/links/5e227e8d92851cafc38c813c/Multinomial-Naive-Bayes-Classification-Model-for-Sentiment-Analysis.pdf

[2] R Albalawi, TH Yeap, and M Benyoucef. 2020. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. Front. *Artif. Intell* 3 (2020), 42. https://doi.org/10.3389/frai.2020.00042

[3] Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. 2020. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence* 3 (2020), 42. https://doi.org/10.3389/frai.2020.00042

[4] Shivam Bansal. [n.d.]. *Beginners Guide to Topic Modeling in Python*. https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/?#.

[5] Diego Buenaño-Fernandez, Mario González, David Gil, and Sergio Luján-Mora. 2020. Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. *IEEE Access* 8 (2020), 35318–35330. https://ieeexplore.ieee.org/iel7/6287639/8948470/09003400.pdf

[6] Yatin Chaudhary, Pankaj Gupta, and Thomas Runkler. 2019. Lifelong Neural Topic Learning in Contextualized Autoregressive Topic Models of Language via Informative Transfers. *arXiv preprint arXiv:1909.13315* (2019).

[7] Tudor Groza, Siegfried Handschuh, and Georgeta Bordea. 2010. Towards automatic extraction of epistemic items from scientific publications. In *Proceedings of the 2010 ACM Symposium on Applied Computing*. 1341–1348.

[8] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information* 10, 4 (2019), 150. https://www.mdpi.com/2078-2489/10/4/150/pdf

[9] Andreas Mueller. [n.d.]. *PyPI WordCloud*. https://pypi.org/project/wordcloud/.

[10] José María González Pinto, Janus Wawrzinek, and Wolf-Tilo Balke. 2019. What Drives Research Efforts? Find Scientific Claims that Count!. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 217–226.

[11] Zhou Tong and Haiyi Zhang. 2016. A text mining research based on LDA topic modelling. In *International Conference on Computer Science, Engineering and Information Technology*. 201–210. https://www.academia.edu/download/45869055/csit65316.pdf

[12] Xin Zhao, Zhe Jiang, and Jeff Gray. 2020. Text Classification and Topic Modeling for Online Discussion Forums: An Empirical Study From the Systems Modeling Community. In *Trends and Applications of Text Summarization Techniques*. IGI Global, 151–186. https://www.researchgate.net/profile/Xin_Zhao136/publication/338302677_Text_Classification_and_Topic_Modeling_for_Online_Discussion_Forums/links/5ec6d95092851c11a87d937e/Text-Classification-and-Topic-Modeling-for-Online-Discussion-Forums.pdf