NATIONAL INSTITUTE OF TECHNOLOGY, SILCHAR

Department of Computer Science and Engineering

Project Report on,

"Examining Trends in Research using LDA Topic Modelling and Trend Prediction"

Submitted by,
Rahul Kumar Gupta (1815125)
Joythish Reddy Evuri (1815079)
Ritu Agarwalla (1815069)
Bukya Hemanth Naik (1815111)
Apil Thapa (1815084)

November 5, 2021

Contents

1 Abstract

Change is the only constant. And in many sectors, we are witnessing a change that is getting increasingly rapid. This change carries opportunities, hazards, and a plethora of new innovation possibilities with it. And this necessitates accurate, well-founded data about potential trends, future developments, and their consequences. Based on this background, this study seeks to catch the main trends, or new directions, paradigms as predictors with an association of each topic which will be discovered through topic modelling techniques like latent Dirichlet allocation (LDA) with N-Grams. For this, we do empirical analysis on 3326 research articles from the journal Applied Intelligence which were gathered during a 30-year span, from 1991 to 2021. The inferred topics are then structured into time series to perform predictive analysis on future trends in research using vector autoregression.
This is significant in the sense that we will be able to predict what technology we will encounter in the future, as well as how far our ability to innovate and discover things may lead us to.

Keywords LDA **.** Apllied intelligence **.** Vector autoregression **.** Topic Modelling

2 Introduction

Research is the key to the advancement of mankind. It is a process of discovering a new domain of knowledge. It is the foundation upon which society progress and advancement are based. Hence it is of utmost importance that the researchers, scholars and all other stakeholders put their resources and energy into the specific area that shows a scope for development. This demands the need for quantitative and predictive analysis of various research trends. For that we need lots and lots of data and the empirical analysis of all those data and a good predictive analysis of all those trends.

A large part of Artificial Intelligence depends upon the information extracted from data. And with the expansion of the internet throughout the world it's tough to get relevant and required information with the expanding volume of data in recent years, most of which is unstructured. However, this has made different technologies take center stage. Topic modelling is one such approach in the field of text classification. It is a process which identifies the topic of a text corpus. The most popular technique is the **Latent Dirichlet Allocation** (LDA). LDA's goal is to determine which topic a document belongs to based on the words it contains.
So the topic modelling techniques can be helpful to study and analyze the underlying topic patterns in different articles and hence can be used to do a quantitative analysis of research trends. Other than that we have different algorithms like **Vector Autoregression** (VAR) which can be used to do a predictive analysis on those trends to help mankind understand the current and future state of different research.

So here, in this research work, we try to do the analysis of research trends in the 3000+ articles published under **"Applied Intelligence"** from 1991 to 2021, using the application of LDA and further predict the future growth of different domains of Applied Intelligence using vector autoregression and other statistical models.

VAR is a statistical model which captures the linear relationship between multiple entities over a time period. So in the later part of the research the results of LDA have been fed into the VAR forecasting model to do predictive analytics on future trends.

3 Methodology

The flow of proceedings for the design of the system has been appropriately described in this section.From the 3326 articles which have been published in the year 1991 to 2021 in the applied intelligence journal.We will thoroughly discuss about data collection in the coming subsection, for now the complete flow diagram where our system design could be depicted is shown below :

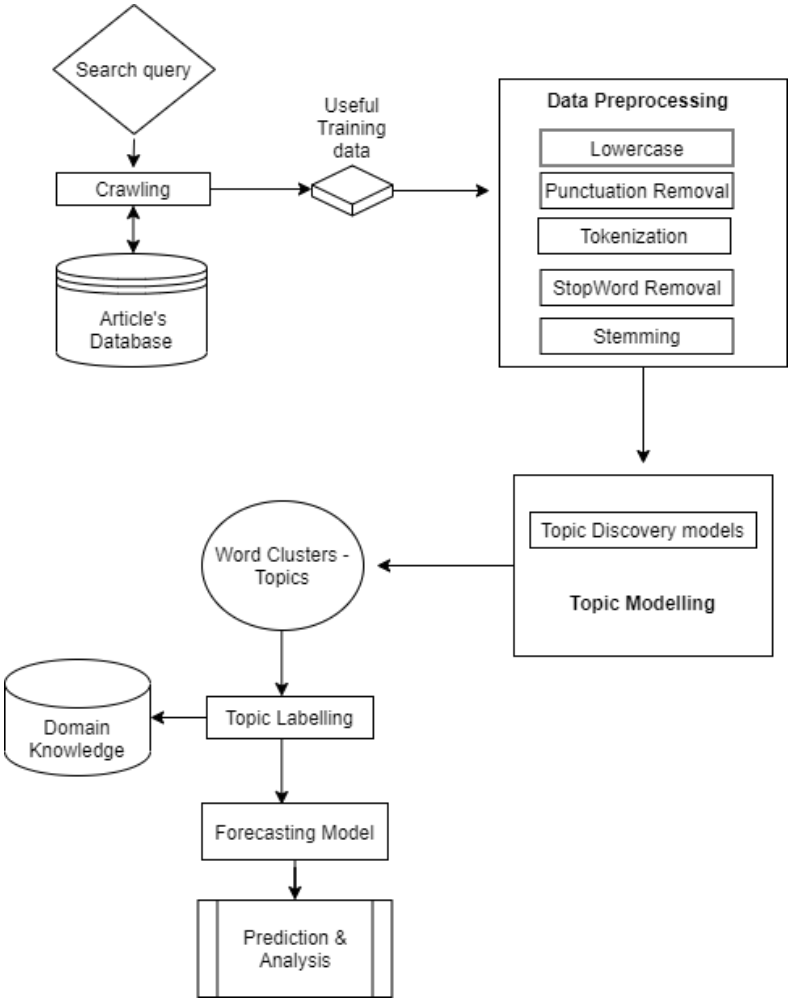

Figure 1: Flow Diagram of system

The four basics steps to be involved are discussed in detail -

3.1 Data Collection

For now, we have collected the articles' useful data viz., date of publication,title and abstract of the article.
For this purpose the technique of web crawling was employed to crawl over the entire database of this journal between the chosen timeline and the extracted details were stored in a CSV(comma separated values) file.

3.2 Pre-processing

An unstructured data usually contains many irrelevant information which we need to get rid of before actually passing on the data for training and analysis.This stage involves many steps, but we are following tokenization,stop-word removal and lemmatization techniques for pre-processing of abstracts of the articles. The above things have been discussed her below :

- **Lowercase** : We will transform our collected data to an uniform case, i.e. lowercase.This makes our further proceedings easier.
- **Punctuation Removal** : There is a need to remove the punctuation because some of the word emebedding models don't suport them.We should carefully choose the list of punctuations which we are going to discard based on the use case.
- **Tokenization** : Tokenization is a way of alienating bits of text into smaller units which are referred to as tokens.This has two ways to be done, word level or sentence level.
- **Stop Word Removal** : These are basically the words that doesn't adds much value to the semantics of the content.After tokenization,this must have to be done which has the following benefits :

    1. Size of the dataset eventually decreases after the removal of stopwords which also leads to reduced training time.
    2. Elimination of these unessentials will help boosting the performance as we have narrowed down the space.Hence, this could enhance the accuracy of classification in the next steps.

- **Stemming** : Stemming reduces the inflected words. It usually refers to a crude heuristic process that get rids of the ends of words in the aim of achieving this goal correctly most of the time, and it often includes the removal of derivational affixes.For example : 'is','am' and 'are' will eventually get converted to 'be'.

After going through above steps, the data will be in the desired form with least unessential words and we can use this as an input to the next step of our system, i.e. Topic Modelling.

4 Results and Discussion

Discussion here

5 Conclusion

Conclusion here

References

[1] S. Sivanandham, A. Sathish Kumar, R. Pradeep, and Rajeswari Sridhar Analysing Research Trends Using Topic Modelling and Trend Prediction

[2] Thoudam Doren Singh, Divyansha,Apoorva Vikram Singh,Abdullah Faiz Ur Rahman Khilji A Hybrid Classification Approach using Topic Modeling and Graph Convolution Networks,North-Eastern Hill University, Shillong, Meghalaya, India. July 2–4, 2020

[3] Canadian Journal of Communication Vol 37 (2012) 189-192 ©2012 Canadian Journal of Communication Corporation

[4] Research Collaboration LLOYD D. FISHER, PHD,* THOMAS ROBERTSON, MD,t G. M. K. HUGHES, PHD.$ARTHURHARTZ, MD, PHD$, PETER LIU, MD,/1 BRUCE F. WALLER, MD,ll MARK YOUNG, MD** Seattle, Washington

[5]

[6]

[7]

[8]

[9]

[10]

[11]

[12] https://medium.com/mlearning-ai/basic-steps-in-natural-language-processing-pipeline-763cd299dd99