Topics Prediction System

Sagar Socrates S, HKBK College of Engineering Bangalore & India, sagarsocrates22@gmail.com

Mohd Usama, HKBK College of Engineering Bangalore & India, MUsama318@gmail.com

Abstract - In today’s world the most pressing matter on a student’s hand is to qualify the examination and most of the times a student does not have clear idea of the topics to refer from the behemoth of a syllabus. In general, usually students approach their teachers/faculties or buy reference books and question banks to minimize their study but they still end up with a huge number of topics to cover for the examination, so the students end up skipping many topics and focuses only on few. In the given paper, we consider a large dataset of almost three years and analyze the dynamics of topical popularity over certain amount of time and considering the various aspects affecting the popularity of a topic or its acceptance in examination papers. We are proposing a model to predict topics based on their frequency to help the students prepare better using the topics suggested.

***Keywords* — Dataset, Examination Papers, Massive Syllabus, Reference books, Topic Popularity, Teachers.**

I. Introduction

“The art of studying and learning consists in remembering the essentials and forgetting what is not.”
- Adolf Hitler, Mein Kampf

In recent years, Predictive analysis using machine learning had become very popular to provide facilities in various fields. In this paper, we plan to use classification of text to tabulate previous questions asked in the examination and predict the most possible topicthat can be asked in coming future. Text classification has always been a critical application and can be used over any textual document.

Predominantly, text classification incorporates text genre-based classification and topic-based text classification. Texts can be written in many genres, like scientific articles, examination papers and so on. Topic-based text classification categorizes documents according to their topics. The main objective of text classification is to allow users to retrieve valuable information from textual resources and deals with multiple operations like retrieval, classification of supervised, unsupervised and semi supervised and summarization of information. Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to classify and discover patterns from textual data.

In the given study, we plan to analyze the topic popularity in examination papers over time. In other words, how the exam papers' knowledge baseis changing overtime with the inclusion of new topics. The primary interest of this study is to identify the patterns in the question paper and growth of popularity of the question topics to predict the topics with the highest probability for the students to study. Comprehending the frequency of topics is necessary because it helps us recognize the patterns in question paper. This study has a direct application in recommendation of the topics to various students to cover massive syllabusin short amount of time.

II. Existing System

There is various Topic Modeling system which makes use text classification and NLP to extract information from the textual document.

[1] This paper reviews various classification techniques that can be used to classify a text file but it does not explore the possibility of suggesting important topics from the classified texts.

[6] This paper the author describes the detailed experimental framework. Their objective is when some features for topical popularity is given, to predict the highest frequency topics at certain point. They take question popularity in account along with other factors

[12] This paper the author examines the prediction of stock using text classification of economic news by analyzing various textual representation.

III. Proposed System

The proposed application aims to suggest topics to the user using text classification on several previously asked questions on a particular subject using topic modeling with the help of various python modules available and also using the semantics of Natural Language Processing and LDA to breakdown the text into smaller understandable pieces which uses Machine Learning. Data sets are created manually by us and are question papers asked in previous years.

DOI : 10.35291/2454-9150.2022.0415

IV. METHODOLOGY

i) Problem Statement-

Text Classification uses Topic modeling to summarize, categorize and segregate the various digital documents based on their content. Text classification comes in handy when you have a large set of data in a digital textual form. Fora student to study more efficiently this project uses the aforementioned algorithms and a little machine learning to extract various topics asked in previous question papers and suggest topics for them to prepare for upcoming examination. Figure 1 shows the working of the proposed model.

ii) Implementation-

The general model works in 3 phases:

a) Data Processing

b) Classification

c) Suggesting the Topics as output

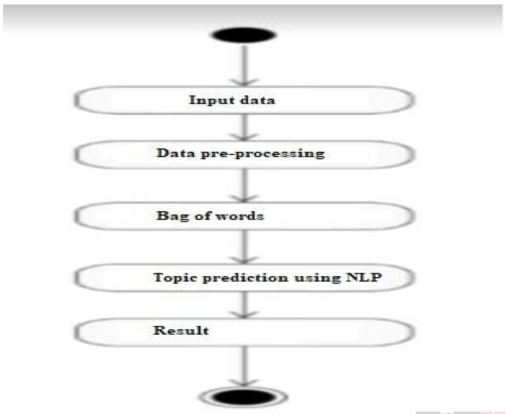


Figure 1: System Architecture

a) Data Processing- comes with a preprocessing module such as Spacy, lemmatization and NLTK for just that purpose. The process string method makes it very easy to prepare text. By default, it strips punctuation, HTML tags, multiple white spaces, non-alphabetic characters, and stop words, and even stems the text, It this provides a processed string. The data is divided into training and testing dataset based on the task faced, it is not essential that 70% of the data has to be for training and rest for testing.

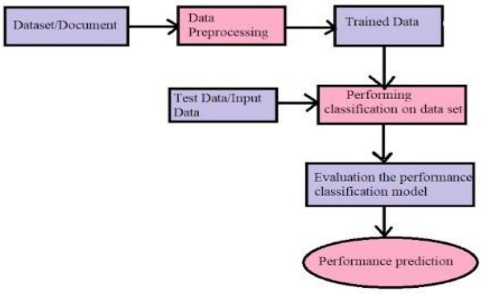


Figure 2: Data Processing

b) Classification- The input from the dataset or input from the user will be a simple stream of Unicode characters such as UTF-8, the classification part requires to convert this given stream to conventional lexical terms like words, phrases, and syntactic markers which can be used to better comprehend the content.

The first step in this is word tokenization which is the process to split the text into a list of words. Basis Technology offers a fully featured language identification and text analytics package (called Rosette Base Linguistics) which is often a good first step to any language processing software. It contains language identification, tokenization, sentence detection, lemmatization, decompounding, and noun phrase extraction.

Lemmatization/Stemming – reduces word variations to simpler forms that may help increase the coverage of NLP utilities.

- Lemmatization uses a language dictionary to perform an accurate reduction to root words.

- Stemming uses simple pattern matching to simply remove suffixes of tokens (e.g. remove “s”, remove“ing”, etc.).

Phrase extraction – extracts sequences of tokens (phrases) that have a strong meaning which is independent of the words when treated separately. These sequences should be treated as a single unit when doing NLP.

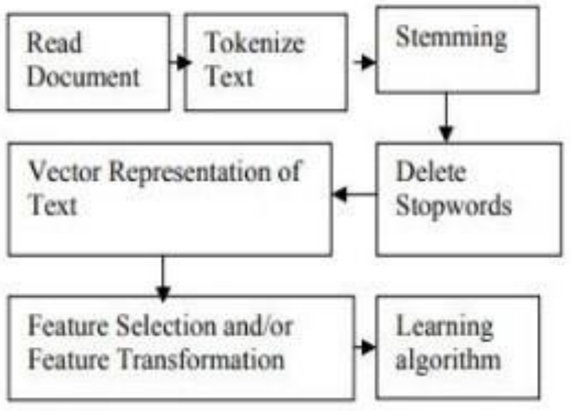

Read Document
Tokenize Text
Stemming
Vector Representation of Text
Delete Stopwords
Feature Selection and/or Feature Transformation
Learning algorithm

Figure 3: Simple Classification Architecture

c) Output- The final stage is to take the classified text and by using libraries like, NLTK, Gensim, spaCy extract the topics and suggest it to the User. Libraries used and Data Flow Diagrams are explained below.

NLTK- It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

DOI : 10.35291/2454-9150.2022.0415

Gensim- Gensim is a Python library for topic modelling, document indexing and similarity retrieval with large corpora. Extractive Text Summarization using Gensim Summarization is a useful tool for varied textual applications that aims to highlight important information within a large corpus.

SpaCy- provides an exceptionally efficient statistical system for NLP in python, which can assign labels to groups of tokens which arecontiguous.

Data Flow Diagrams-

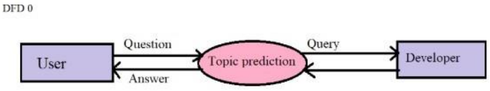


Figure 4: DFD Level 0

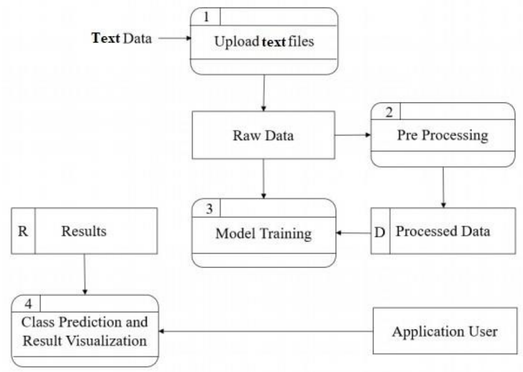


Figure 5: DFD Level 1

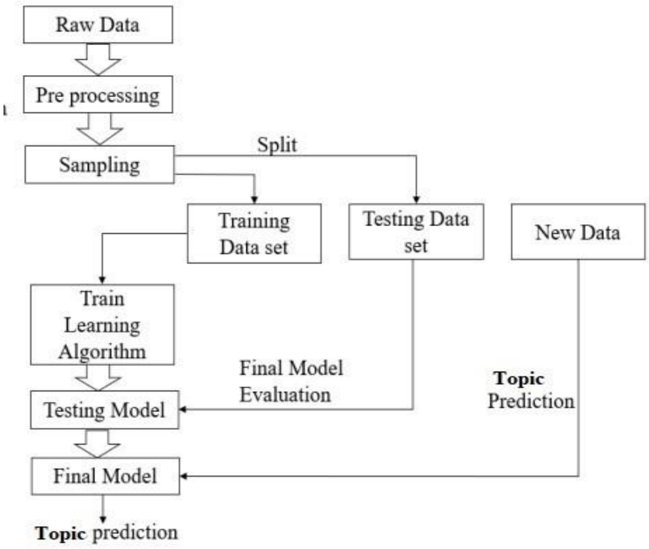


Figure 6: DFD Level 2

V. Results And Discussion

Many tests were conducted to check the integrity of the system, it is proved that the system works as expected and provides the expected output as topics from the given dataset which are to be suggested to the user. These tests included different unit tests and integration tests.

Use case diagram for final resulted system is given below:

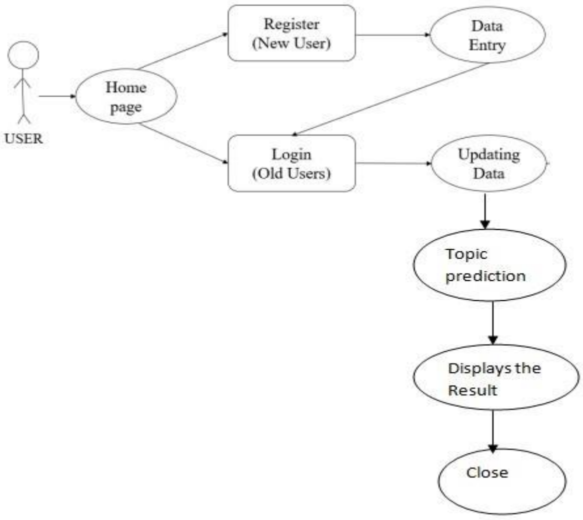
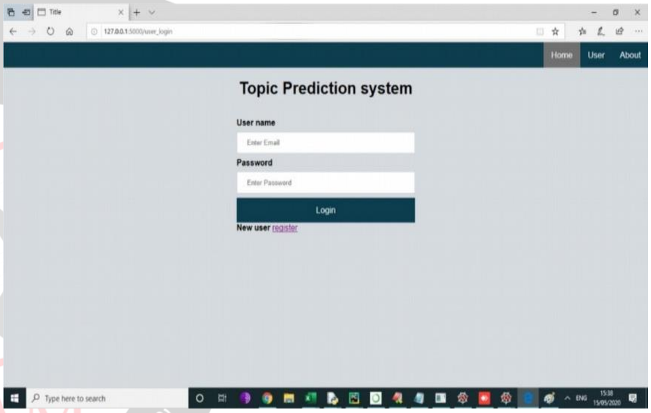


Figure 7: Use Case Diagram

Whenever a user accesses the application, he/she will be presented with a login page so that many users can create their profile and get topics suggested to them from different subjects. Each user will have their own login ID or they can create one with the option available



Topic Prediction system

User name

Enter Email

Password

Enter Password

Login

New user register

Figure 8: Login Page

And finally, Users can upload a csv file as dataset which should have two fields Question number and Question, this dataset will be processed and topics will be suggested to the user.

Optionally there is a field provided to upload another file say Syllabus copy and the model will compare both syllabus copy and dataset to provide topics which are not available in either of the copy, this way user can compare their suggested topic with the syllabus provided to them.

DOI : 10.35291/2454-9150.2022.0415

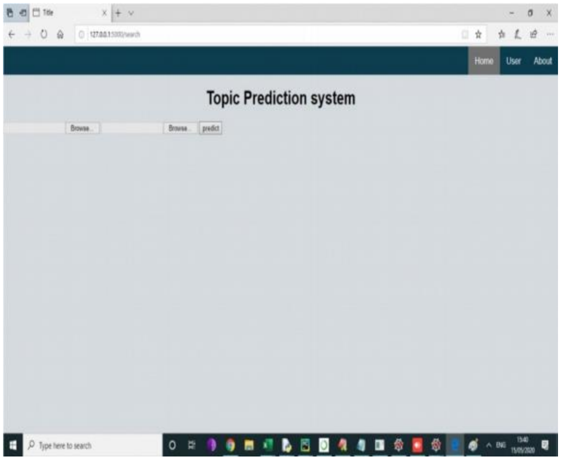


Figure 9: Output page 1

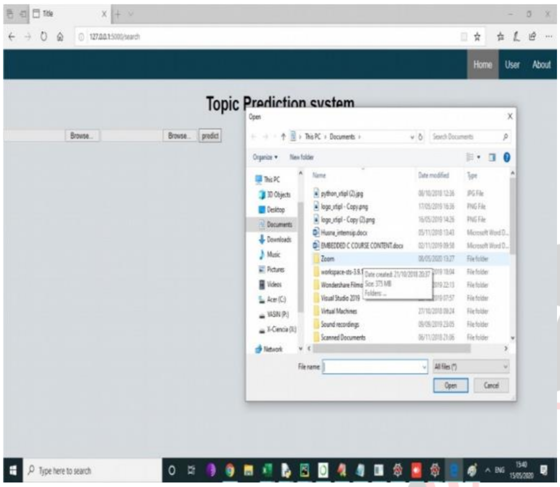
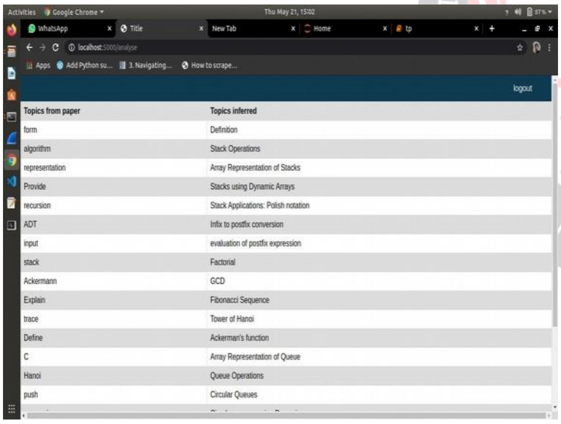


Figure 10: Output page 2- Selecting the file



Topics from paper	Topics inferred
form	Definition
algorithm	Stack Operations
representation	Array Representation of Stacks
Provide	Stacks using Dynamic Arrays
recursion	Stack Applications: Polish notation
ADT	Infix to postfix conversion
input	evaluation of postfix expression
stack	Factorial
Ackermann	GCD
Explain	Fibonacci Sequence
trace	Tower of Hanoi
Define	Ackerman's function
C	Array Representation of Queue
Hanoi	Queue Operations
push	Circular Queues

Figure 11: Output page 3- Extracted Topics

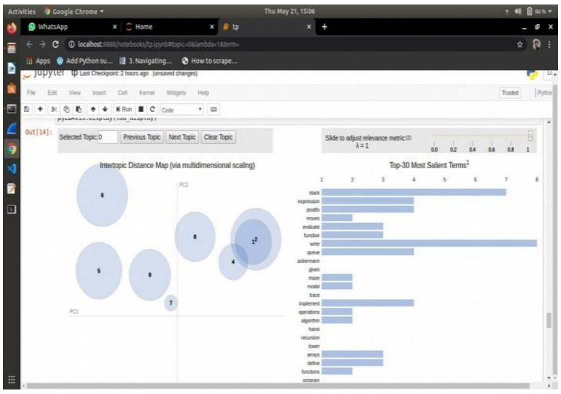

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:
λ = 1
Intertopic Distance Map (via multidimensional scaling)
Top-30 Most Salient Terms
PC1
PC2

Figure 12: Output 3- Visualization based on Topic Weights

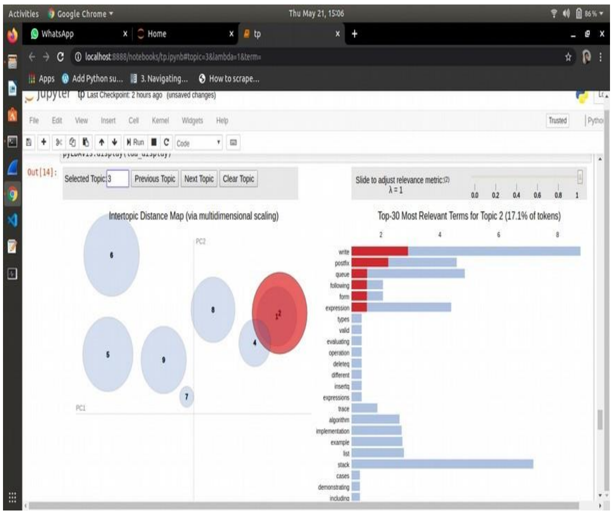


Figure 13: Output 4- Visualization of 3rd topic w.r.t others

VI. Conclusion

The main aim of this paper was to provide important topics to the students by taking input as question paper asked in previous examination, we have been successful in implementing the Topics Prediction System Based on the keywords that are derived from the entered questions, various cognitive levels are chosen. The choice of Topic prediction mapping is then based on the levels that have been chosen. In order to execute outcome-based predictive analysis, classification algorithms are applied to the topic prediction attainment data. We started by using statistical analysis to investigate the variables that affected the task of predicting question popularity. After that, we suggested using supervised machine learning to model the influential aspects. So, in order to predict the topics that students should study, we developed a system that can take text classification and calculate the dynamics of topical growth and popularity over time. It should be feasible, generates relevant topics with highest probability of them to occur in examination and should provide accurate results.

VII. Future Scope

This version of the project takes the basic approach on topic modeling and does not dwell deep on machine learning implementation to make much more precise and accurate prediction of topics. Many Deep Learning techniques can be incorporated as a future improvement in the project to make it more amusing like, instead of using just topic modeling on the dataset one can provide a list of topics beforehand and the processed dataset will be classified according to the given topics using LDA, weights can be assigned tothe topics for suggestion. This process can also be achieved by changing unsupervised LDA to semi supervised LDA also called *GuidedLDA*. GuidedLDA uses probability to classify the document by each word.

DOI : 10.35291/2454-9150.2022.0415

Another implementation that can be introduced is Linguistic Models, they are unsupervised learning techniques creates RNN embedding on large texts corpora to understand the language of the texts before training the model for classification. Text vectorization can be improved with methods like *word2vec* and *tfidf* and Parts of speech tagging can be used with taggers like *NLTK POS tagger* and *Stanford POS tagger*. Classification can be performed using LSTM by passing vectorized words to extract the topics from the processed document.

REFERENCES

[1] Ikonomakis, Emmanouil & Kotsiantis, Sotiris & Tampakas, V.. (2005). Text Classification Using Machine Learning Techniques. WSEAS transactions on computers. 4. 966-974.

[2] Kim S. B., Rim H. C., Yook D. S. and Lim H. S., “Effective Methods for Improving Naive Bayes Text Classifiers”, LNAI 2417, 2002, pp. 414-423

[3] Schneider KM. (2005) Techniques for Improving the Performance of Naive Bayes for Text Classification. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science, vol 3406. Springer, Berlin, Heidelberg

[4] Klopotek M. and Woch M., “Very Large Bayesian Networks in Text Classification”, ICCS 2003, LNCS 2657, 2003, pp. 397- 406

[5] Shanahan J. and Roma N., Improving SVM Text Classification Performance through Threshold Adjustment, LNAI 2837, 2003, 361-372

[6] Analysis and Prediction of Question Topic Popularity in Community Q&A Sites: A Case Study of Quora Suman Kalyan Maity, Jot Sarup Singh Sahni and Animesh Mukherjee

[7] [7] Ma, Z.; Sun, A.; and Cong, G. 2012. SIGIR ’12, 1173–1174. New York, NY,USA: ACM.

[8] C .C. Aggarwal and C.X. Zhai(eds.),Mining Text Data, DOI 10.1007/978-1- 4614-3223-4_6, Springer Science+Business Media, LLC 2012.

[9] B. Liu, L. Zhang. A Survey of Opinion Mining and Sentiment Analysis. Book Chapter in Mining Text Data, Ed. C. Aggarwal, C. Zhai, Springer, 2011.

[10] S. Chakrabarti, B. Dom. R. Agrawal, P. Raghavan. Using taxonomy, discriminants and signatures for navigating in text databases, VLDB Conference, 1997.

[11] A. Y. Ng, M. I. Jordan. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. NIPS. pp. 841848, 2001.

[12] Schumaker, R. P. and Chen, H. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans.Inform.Syst.27, 2, Article 12 (February2009), 19 pages.

[13] Moldovan, D., Paşca, M., Harabagiu, S., & Surdeanu, M. (2003). Performance issues and error analysis in an open- domain question answering system. ACM Transactions on Information Systems (TOIS), 21(2), 133-154.

[14] Tolle, K. M., & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American society for information science, 51(4), 352-370.

[15] Sekine, S., & Nobata, C. (2004, May). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In LREC (pp. 1977-1980).

DOI : 10.35291/2454-9150.2022.0415