

**School of Computer Science and Engineering (SCOPE)**

**Btech- Computer Science and Engineering**

**CSE3013 – Artificial Intelligence**

**J Component**

**Review-1**

Title: Indexing Contents Of Documents Using LDA

Submitted By

Jaivant Vassan – 19BCE2322

Kabhilan S – 19BCE2339

Prithvi Saran S – 19BCE2344

Faculty: Dr. Geraldine Bessie Amali D

Slot: C2

**Problem Formulation:**

Majority of all data available is unstructured and unlabelled data. Extensive progress has been made in the field of text processing to bring some semblance of order to these data for ease of information retrieval and processing. Topic modelling and topic classification are two such areas of research. While topic modelling has always been less accurate in comparison, topic modelling based models do not require labelled data for training, which is a huge advantage.

We plan to design an API that is capable of annotating unlabelled data in documents, by collectively analysing sets of documents using an unsupervised topic modelling algorithm, LDA(Latent Dirichlet Allocation) and identifying the focal points of interests discussed in the documents.

**Literature Survey:**

**[1] A Text Mining Research Based On LDA Topic Modelling**

The researchers Zhou Tong and Haiyi Zhang reported on a text mining research based on LDA topic modelling, which could be used for social and business research. The research is intended to design a new computational tool based on topic models using text mining techniques. By applying Latent Dirichlet allocation (LDA) and topic modelling, a solution of topic searching, exploring and recommending system will be achieved. Text mining is the process of deriving high-quality information from text. The researchers evaluated 200000 articles taken from Wikipedia and tweets from 100 twitter users with more than 100 tweets each. Results showed that they were able to accurately categorise articles and a user’s tweets into separate topics according to probability of use of each topic. They recommend that with these data and model foundation, a number of future works can be done for further research and experiment.

**[2] Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data**

In this paper published by Ireneus Kagashe, Zhijun Yan , and Imran Suheryani, the dataset was extracted from twitter using specific parameters with the python package tweepy. Although they’ve used an LDA model to extract the survey info, in order to achieve accurate results, they had to manually filter out the unrelated tweets. Also during feature extraction, they identified the words directly dependent on the name of the drugs, assigned them weights using the tf-idf measure. Thus helping the SVM model(the model they used for training )to identify widely used drugs as it would now have better features opposed to the plain text before. Although this tremendously increases the accuracy of the model, when large amount of data has to be processed this method would prove inefficient.

**[3] Report on Text Classification using CNN, RNN & HAN**

This report by Akshat Maheshwari, provided great insight into the performance of Recurrent and Convoluted neural networks when used to classify text. Preprocessing: the stop words and other unwanted tags and symbols are removed using the python package beautifulsoup.The data is cleaned using the same method for all three neural network models. The CNN model performs very well in identifying n-grams(words stringed together) but it does so without taking context into account i.e., the model’s layers process an entry and the activation function adds to the weight only when a similar word group from previous data is encountered. This could lead to inconsequential word patterns getting more priority. In RNNs, the preprocessing also includes one-hot encoding using countvectorizer from the sci-kit learn package or the keras package. RNNs keep track of long chains of recursions thorught their layers, and will therefore be useful to catch long term dependencies. During training, bidirectional LSTM layers are used to encode all the weight assignment info forward into the last layer(output).

**[4] Analysing the Social-Economic Impact of Wireless Mobile Services During and Before COVID-19 Using Topic Modelling and Sentiment Analysis on Tweets**

In this article by Wehel Hadi, three models of analysis are compared by implementing them to measure the impact of wireless mobile services on customer sentiments like user happiness (social effect), affordability (economic effect), and willingness (social effect). Twitter is used as the test environment where user reactions are similar to that in a general social setting. This investigation is also based on two different mobile phone companies with differing flexibilities in their policy to investigate whether a company’s policy influences the sentiment of core values such as user satisfaction (social effect), affordability (economic effect) and willingness (social effect). In conclusion, theres a negative customer sentiment especially higher due to the pandemic period. The mobile company with a more flexible policy is observed to have lesser negative rate in comparison to a company with more traditional policies. In analysis using the software called PyLDAvis using several classifiers and being optimized using hyper parameter tuning, over four social topics, LDA proves to be an excellent model with 98% accuracy and the highest coherence rate of 0.55 in comparison to LSA with 0.52 and HDP with 0.19.

**[5] More Than Words: Collocation Tokenization for Latent Dirichlet Allocation Models**

In this paper, the shortcoming of LDA that is to achieve the best results for languages without marked word boundaries such as Chinese and Thai is rectified by using the Pearson’s chi-squared, t statistics and Word Pair Encoding (WPE) to produce tokens as input to the LDA model. These input tokens are trained on wikipedia text to look for the different word types that should be grouped together such as compound nouns, proper nouns, and complex event verbs. A new metric is proposed for measuring the clustering quality where the vocabularies of the models differ significantly. Based on various metrics like Held-Out Likelihood & Concatenation-based Embedding Silhouette (CBES) and other established metrics, it is established that the topics trained with merged tokens result in topic keys that are clearer, more coherent, and more effective at distinguishing topics than those unmerged models.

**[6] Machine learning in medicine: a practical introduction to natural language processing**

In this research article by Conrad J.Harrison, publicly available data from medicine review websites are used to conduct three NLP experiments namely lexicon-based sentiment analysis on open-text patient reviews of four drugs: levothyroxine, Viagra, oseltamivir, and apixaban followed by unsupervised ML (latent Dirichlet allocation, LDA) to discover related medications in the dataset based simply on their reviews. Three supervised machine learning algorithms to determine whether a drug review was good or negative is made. On comparing the results of various algorithms, LDA proves to be an efficient model in providing replicable code that may be easily adapted to different research investigations using open-source tools, as well as a conceptual overview of typical methodologies used to analyze enormous quantities of text such as the medical data used in this article.

**[7] A Semi-Supervised Bayesian Network Model for Microblog Topic Classification**

A team of researchers have proposed a novel way of classifying messages on microblogging services. The team led by Yan hen from the State Key Laboratory of Software Development Environment noted that the approach is to submit a query that is related to hashtag and category to Google Search Engine. It incorporates the external information provided by search engine results to enrich the short microblogs. It is interesting to explore whether the incorporation of social network structure can improve the performance of microblogging classification. Microblogging services are becoming immensely popular in breaking-news disseminating, information sharing, and events participation. They proposed a novel scheme to classify microblogging messages, which addresses three concerns in microblog classifications. The semi-supervised classifier seamlessly fuse labeled data structure and external resources into the training process. The authors reviewed 20 extended documents. The incorporation of external resources to supplement the short microblogs well compensates the data sparseness issue.