Analysing Retail Sector News using NLP during COVID-19

Amani Goli

2944590

Submitted in partial fulfillment for the degree of

Master of Science in Big Data Management and Analytics

Griffith College Dublin

September 2020

Under the supervision of **Dr Aqeel Kazmi**

Disclaimer

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Applied Digital Media at Griffith College Dublin, is entirely my own work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Signed:  **Date: 07/09/2020**

Acknowledgments

First, I would like to thank my beloved Institution ***Griffith College*** for providing this opportunity to explore and practice my knowledge arena in Big Data Management and Analytics.

I owe my sincere gratitude and thanks to my supervisor **Dr Aqeel Kazmi** for his help, advice, and patience during this thesis. He was always really helpful, informative and I learnt a lot from him. The achievements in this study would not be possible without his supervision. I also would like to acknowledge my friends and seniors support who guided me with the knowledge they possess. I would like to extend my warm gratitude towards my family and well-wishers who have been great support morally and made me achieve and reach this stage.

Table of Contents

List of Equations

List of Tables

LIST OF FIGURES

Abstract

We live in an information era. We hear the news everywhere and see it. We open our social media account and our newsfeed contains news. We turn on your television and a news flash is aired. Newspapers here and there being sold out.

Let us talk about the news on Google. How often does one check Google news individually? Every nanosecond or less, it refreshes.

The unstructured text represents 80 percent of the mass of data that flows into information networks and becomes the most common on-line data stored. Large-scale NLP tools are urgently needed to transform this enormous amount of data into readily available information. The problem of extracting novel information from very large unstructured collections of text documents (text data mining) has, in particular, attracted considerable attention. One move towards solving this problem is to arrange the documents into manageable categories according to their content and to imagine the collection, offering an overview of the variety of documents and their relationships to make them easier to search.

Chapter 1. Introduction

1.1 What is Topic Modelling

Topic modelling is an unsupervised technique which aims to analyze large volumes of text data by grouping the documents into groups. In the case of topic modelling, no labels are attached to the text data. Modelling the subject attempts instead to organize the documents into clusters based on common characteristics[1].

A common example of topic modelling is where a large number of newspaper papers belonging to the same genre are grouped together. In other words, cluster documents which have the same subject matter. It is important to note here that the performance of topic modelling is extremely difficult to assess as there are no correct answers. It is up to the user to identify similar characteristics within one cluster 's documents and to assign an acceptable label or subject to it.

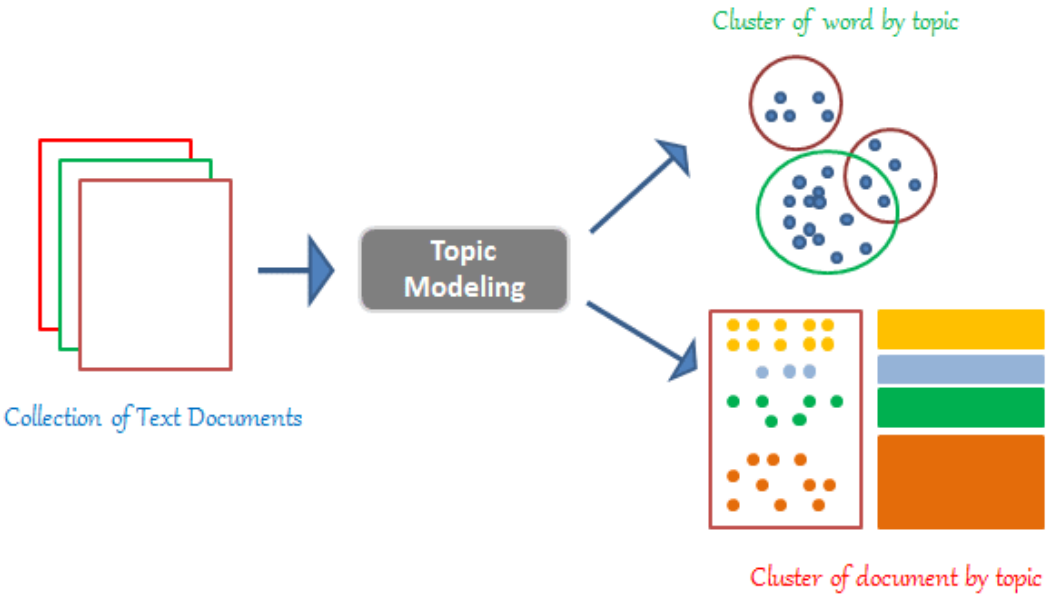


Fig 1.1 Topic Modelling

Why do we need topic modelling?

In an unstructured environment, we can see a large amount of textual data lying around us in the form of news articles, research papers, social media posts etc. and we need a way of recognizing, organizing and marking this data to make informed decisions[1].

Topic modelling is used in various applications, such as identifying stack overflow questions identical to one another, aggregating and analyzing news flows, recommending systems, etc. All of these concentrate on finding the secret thematic structure in the document, as any document we write, be it a tweet, post or research paper, is believed to consist of topics such as sports, physics, aerospace etc.

How to do topic modelling?

There are currently several ways to model topics but we will address a probabilistic modeling method called the Latent Dirichlet Allocation (LDA) developed by Prof. David M. Blei in 2003. This is an extension of the Probabilistic Latent Semantic Analysis (PLSA) introduced by Thomas Hoffman in 1999, with a very minute difference in the way they handle distribution per text[1].

1.2 PRIMARY GOALS

APIs are now live! They are open to regular updates. Let 's talk about the news at Google. How much does one search news from Google. It updates per nanosecond or less. This paper analyzes for any spot, retail news that appears instantly in Google. So let's look at what the Retail industry cooks here.

We want to analyse what news related for a particular region are coming out on a daily basis.

What are the patterns over at the end of each day?

What kind of news which we hear?

Are there any similarity between the news articles?

Is there any pattern between these articles?

And how does this news effect on that particular industry.

We will be using NLP Algorithms, Parametric comparisons. So we will compare with all these different packages and compare the outcome.

What percent of views are coming out of this category. How much Negativity or Positivity coming out for this industry.

1.3 OVERVIEW OF APPROACH

- Extract the dataset
- Extract the data from json format to .csv.
- Pre-process the converted data.
- Exploratory analysis.
- Preparing the LDA analytical results.
- Working on model LDA.
- Evaluating the effects of an LDA model.

1.4 RESEARCH QUESTIONS

- What topics each document (or) news article belong to?
- What are the areas of improvement for any particular sector if identified?
- How many market clusters can be formed from the kind of news we aimed at?
- What category of customers are we focusing from the kind of news we fetched?

Chapter 2. Background

2.1 LITERATURE REVIEW:

The literature review part of the research work helps us to understand the related work that has been already done on the research work that we have chosen for implementation. This phase is considered to be one of the major part, as we can get to know the different approaches that can to be followed and on top of that, it will helps us to anticipate any issue that we may face in the implementation well in advance by analysing the different research works.

2.2 RELATED WORK

Latent Dirichlet allocation can be used in conjunction with machine learning algorithms to identify text documents as a function representation process. Latent Dirichlet allocation (LDA) is a popular approach for uncovering latent topics: multinomial distributions of probability over words, generated by soft word clustering centered on the fact that records co-occur. Although LDA does produce sensitive topics a prominent issue is the existence of \junk topics which include incoherent or insignificant groupings of words. Model outputs are also expected to be validated and updated by domain experts to ensure they adhere to practical theoretical concepts. The related work on representation of features based on LDA is briefly described here.

Hall applied LDA over 14,000 publications to examine research patterns in computational linguistics. The authors were hiring experts to verify the latent rating
Themes. Out of hundred topics thirty six were retained and extra ten topics were added to the model did not produce manually. Talley et al. have taken a gander at 110,000 NIH concedes and distinguished 700 idle subjects utilizing LDA.[2]
A generous measure of modification was remembered for the demonstrating cycle: refreshing the jargon to incorporate abbreviations and multi-word words, killing insignificant subjects, searching for boundaries and looking at the subsequent renditions.

Current effective quality appraisals depend vigorously on the specialists surveying arrangements of the most probable terms in a subject. For instance, in organic writings one may locate a subject with terms "dna, replication, rna, fix, complex, association" The earlier work in representation recommends some elective methods of introduction. Communication may then empower clients to investigate elective orders. An appropriate model of words (e.g., measurably noteworthy instead of continuous terms, expresses as opposed to words) can additionally aid correlation. Joining word relatedness into a representation, elevated level examples can surface in the content. Not at all like current LDA model execution audit apparatuses, Termite plans to help the area explicit assignment of creating and refining subject models.

Tian used the latent method of allocating Dirichlet to index and evaluate the source code documents as a mixture of probabilistic subject matter. Based on this description, the software systems are automatically classified in open-source repositories. Taşçı and Güngör analyzed the success in text categorization of the latent Dirichlet allotment dependent representation. To address the problem of text mining problems with high dimensionality, feature selection methods such as knowledge gain, chi-square statistics, and document frequency threshold are considered. LDA 's output is contrasted with those methods of selecting functions. TF-IDF scheme is used in the comparative assessment to evaluate the terms, and support vector machines are used as the base learners. Ramage et al.[14] presented a labelled latent allocation model for Dirichlet which integrates labels and subject priors to learn word-tag correspondence. The experimental results indicate that, owing to its explicit modelling of the value of each label in the paper, labelled LDA method can yield better output than support vector machine classifier.

Hong and Davison used two thematic modelling approaches (the author-topic model and the latent allocation of Dirichlet) to forecast common Twitter messages and classify Twitter users and related messages into topical categories.

Liu empirically assessed vector space model efficiency, latent semantic indexing, and latent Dirichlet allocation methods on text classification. In the classification of text based on LDA, the latent method of allocating Dirichlet was used to represent the text documents.The experimental results showed that the use of LDA in conjunction to help vector machines produces better performance than the other configurations compared.

Newman (2010) define a tool for rating terms in topics to assist in interpretability, called the rating of pointwise shared knowledge (PMI).Under the PMI ranking of words, in some broad, external "reference" corpus, such as Wikipedia or Google n-grams, each of the ten most likely words within a topic is listed in a decreasing order of approximately how often they occur in close proximity to the nine other most likely terms from that topic.While this approach was highly correlated within topics with human assessments of term significance, it does not easily generalize to topic models suitable for companies that do not have an established source of word co-occurrences readily accessible.

In comparison, Taddy (2011) uses an intrinsic measure to rank terms within topics: a quantity called lift, defined as the ratio of probability of a term within a topic to its corpus-wide marginal probability.This usually reduces the rankings of commonly used global words, which can be helpful. However, we find it can be noisy by, for instance, giving high rankings to very rare terms that occur in just one subject. Although those words may contain useful contextual material, the subject can remain difficult to understand if they are very unusual.

At long last, Bischof (2012) are creating and executing another measurable theme model that deduces both the recurrence of a word and its restrictiveness to the degree that its events are restricted to just a couple of subjects. They execute a univariate measure called a FREX score ("Frequency and EXclusivity") which is a weighted symphonious mean of the position of a word inside a given point regarding recurrence and eliteness, and they propose it as an approach to rate terms to aid the investigation of topic. We propose a comparative cycle, which is a weighted normal of the logarithms of probability of a term and its lifting, and we legitimize it with a client investigation and coordinate it into our intelligent representation.

Lately a scope of perception frameworks have been created for the point models. A considerable lot of them focus on permitting clients to look through papers, subjects, and words to find out about the connections between these three standard model units (Gardner et al., 2010; Chaney and Blei, 2012; Snyder et al., 2013). As a rule, these programs use arrangements of the most potential words inside themes to sum up the points, and the representation components are limited to bar graphs or word billows of term probabilities for each subject, pie diagrams of subject probabilities for every content, as well as various bar outlines or scatterplots applicable to metadata. While these devices can be helpful for looking through a corpus, we are searching for a more com-settlement representation, with the

smaller accentuation on understanding the individual points themselves (without essentially envisioning records) rapidly and without any problem.

Chuang (2012b) make such a device , called "Termite," that utilizes a lattice format to imagine the arrangement of topical term dispersions assessed in LDA. The creators present two measurements of the utility of words to depict a theme model: unmistakable and saliential. These amounts compute how much data a term communicates about points by estimating the Kullback-Liebler disparity between the appropriation of themes given the term and the negligible circulation of subjects (peculiarity), ideally weighted by the general recurrence (saliency) of the word. As a thresholding instrument for picking the words are remembered for the graph, the creators recommend saliency, and they additionally utilize a seriation framework to arrange the most well known terms to show errors between subjects.

Termite is a lightweight, instinctive intuitive representation of the themes in a point model, yet it is restricted to offering a worldwide perspective on the model as opposed to empowering a client to inspect explicit points in detail by imagining a hypothetically extraordinary arrangement of terms for each subject. Chuang et al. (2013a) explicitly characterize the utilization of a "point explicit word requesting" as conceivably helpful future work.

Another exploration bunch focussed on programming designing subject displaying, Eth et al. They utilized LDA just because to remove and perform subjects in source code. PC likeness perception. That is, LDA utilizes an instinctive way to deal with Calculation of the similitude between source records with the particular appropriations of and source Document subjects done. They have utilized their methodology on 1,555 Apache programming ventures. What's more, Source Forge which contains 19 million lines of code (SLOC). The creators have exhibited this strategy, can be effective for venture association, refactoring programming .

Tian actualized a computerized LDA-based strategy for PC structure categorisation, named LACT. Utilizing 43 open hotspot for LACT evaluation The product frameworks can be arranged in various programming dialects and demonstrated LACT Of programming language-based programming frameworks (Tian et al., 2009). Lukinet et al. Proposed a LDA-based displaying way to deal with the subject for bug use Placeback. Their idea was stretched out to the investigation and impacts of similar bugs in Mozilla and Eclipse

Indicated that their LDA-based way to deal with finding and inspecting bugs is better than LSI .Such code causes (Lukins et al., 2008, Lukins et al., 2010).

Z.Zhai use related knowledge in the LDA models as a limitation to upgrade gathering of attributes LDA has. They should connection and they can't join limitation from the corpus. Must connection infers that there must be two highlights in a similar classification while can't interface limits two can't have highlights in a similar classification. These requirements are instantly suspended. Provided that this is true, at that point at any rate one of the implications of two item includes is the equivalent, and is thought to be a similar network as must be connected. By correlation, if two highlights are introduced in a similar expression, they are viewed as a different capacity without the "and" combination, and ought to be in different classes, as they can't interface (Zhai et al., 2011).

Wang have proposed a LDA-based methodology that can be called Bio-LDA Identify the natural jargon used to secure inert topic. The scholars revealed that approach can be actualized in different examinations, for example, journey for affiliations, affiliations Predication, and formation of availability maps. What's more, they have indicated that Bio-LDA can be applied to build the utilization of sub-atomic holding strategies as warmth maps (Wang et al., 2011).

Fang proposed another non-administered LDA-based subject model for contrastive sentiment demonstrating planned for looking for feelings from alternate points of view dependent on a given theme and their distinction on the point with qualifying rules, the model called Cross-Perspective Topic (CPT) model. They led tests for both subjective and quantitative estimations on two political datasets that include: first dataset is affirmation records of U.S. legislators indicating representatives' political situations through these archives, and for the second dataset, got from globe. News Medias in the U.S. (New York Times), China (Xinhua News) and India (The Hindu) from three delegate media. Utilizing corrIDA and LDA as two baselines for contrasting their methodology and different models (Fang et al., 2012).

Yano utilized a few LDA-based probabilistic models to foresee reactions from political blog entries. In more detail, they utilized theme models LinkLDA and CommentLDA to create blog data(topics, post words) in their techniques, and with this model a relationship can be found between the post, the analysts and their replies.To test, their model gathered remarks and blog entries from 40 site pages with an accentuation on American governmental issues (Yano et al., 2009, Yano and Smith , 2010).

Madan Introduced another strategy for general detecting zeroed in on the utilization of phone sensors and utilized a LDA subject model to find patterns and investigation of individuals' practices that changed their political feelings, even assessed diverse political conclusions for singular residents, considered a proportion of homophilia elements that uncovers patterns for outside political occasions. They got a phone to accumulate information and apply their methodology detecting device for recording social elements and unexpected factors from John McCain's and Barack Obama's most recent three months of 2008 American Presidential missions (Madan et al., 2011). They broke down enthusiastic responses from Balasubramanian et al. An epic model Multi Group Response LDA (MCR-LDA) was proposed, which is in actuality a multi-target and utilized sLDA for anticipating remark extremity from post substance and supporting vector machine order (Balasubramanian et al., 2012). To assess their methodology, they gave a dataset of blog entries from five websites zeroing in on the US strategy made by them (Yano et al., 2009)[11].

Chen recommended a generative model to auto-find the inactive relationship between supposition words and themes that could be helpful for the extraction of political perspectives and utilized a LDA model to diminish the size of modifier words, the creators prevailing with regards to extricating those expressions from their model and indicated that this model could be powerful in various perspectives. They zeroed in on representatives' announcement records which included 15, 512 proclamations from 88 congresspersons on the Project Vote Smart Website (Chen et al., 2010). Tune et al. It was analyzed on Twitter how social and policy driven issues identified with South Korean presidential races in 2012, and utilized a LDA technique to assess the connection among occasions and tweets points (Song et al., 2014)[3].

Zirn proposed, in light of an expansion of LDA, a strategy for assessing and looking at archives and utilized Logic LDA and Labeled LDA approaches in their technique for displaying subjects. Since 1990, they considered German National Elections as a dataset to apply their technique and show that utilizing their strategy is reliably better than utilizing a reference approach that reproduces manual comment dependent on assessment of text and catchphrases (Zirn and Stuckenschmidt, 2014)[9].

Huang built up a LDA-based instrument for the ID of inside treatment designs for clinical cycles (CPs), and the identification of these mystery designs is really one of the most significant components of clinical cycle evaluation. Their key technique is to acquire care

stream logs, and furthermore to appraise concealed patterns dependent on LDA for the gathered logs. Built up patterns will meet all requirements for grouping and find clinical practices that have a similar clinical consideration. Utilizing an informational index acquired from China's Zhejiang Huzhou Central Hospital (Huang et al., 2013) to explore different avenues regarding the capability of their strategy.

Liu, They actualized a model for the disclosure of practical miRNA administrative modules (FMRMs) that consolidate heterogeneous datasets and coordinate articulation profiles of both miRNAs and mRNAs, utilizing or in any event, utilizing the past objective restricting data. This model utilized a subject model dependent on the Latent Dirichlet Allocation of Correspondence (Corr-LDA). They play out their strategy for mouse model articulation informational collections as an evaluation dataset to explore the human bosom disease issue. The creators locate their model successful in getting different naturally important models (Liu et al., 2010). Zhang u.a. The creators had a subject demonstrating investigation on the conclusion of Chinese medication (CM) and built up a model dependent on the Author-Topic model to identify CM analysis from Clinical Knowledge of Diabetes Patients, and named the Symptom-Herb-Diagnosis (SHDT) model. 328 patients with diabetes were gotten from the appraisal dataset. The outcomes showed that in comorbidity ailments, (for example, cardiovascular infection and diabetic kidney) the SHDT model can find natural remedy themes and regular manifestations for a lot of significant clinical related sicknesses (Zhang et al., 2011)[6].

Drape Authors in this work focused on the issue of characterizing bunch literary subjects like spatial antiquities with text portrayals. They proposed crossover strategies dependent on bunch system and point model to find groups of literary articles from geo-found archives. Truly, they utilized a generative probabilistic model (LDA) and the DBSCAN calculation to discover points from reports. They utilized the Reuters-21578 dataset as a dataset in this paper to test their strategies (Zhang et al., 2015).[7]

McInerney introduced an investigation on the characterization of noteworthy Twitter posts, The creators applied a probabilistic model to the subject disclosure in the geographic territory and this model can discover concealed critical occasions on Twitter and furthermore thought to be stochastic variational induction (SVI) to apply angle rising to the vector objective with LDA. They gathered 2,535 geo-labeled tweets from New York 's Upper Manhattan territory. That KL dissimilarity is a decent measurement for recognizing a critical

tweet occasion, however the outcome will be negative for an enormous dataset of news stories (McInerney and Blei, 2014).

Tian presented a LDA-based technique for the programmed categorisation of programming frameworks , called LACT. For LACT assessment, 43 open-source programming frameworks were utilized in different programming dialects and LACT indicated that product frameworks can be sorted dependent on the programming language type (Tian et al., 2009).

Lukinet Proposed a methodology displaying of the subject dependent on LDA model for bug limitation purposes. Their thought, applied to the investigation of similar bugs in Mozilla and Eclipse, demonstrated that their LDA-based way to deal with assessing and breaking down bugs in these source codes is better than LSI (Lukins et al., 2008, Lukins et al., 2010).

Yang by considering the mix of definition and basic information stream information, they executed a subject explicit methodology, utilizing a serious LDA-based theme model with GA to comprehend pernicious applications, bunch applications as per their depictions. They likewise utilized their technique on kind application 3691 and threatening application 1612. In featuring pernicious activities, the creators discovered point explicit, information stream marks are amazing and helpful (Yang et al.,2017).

Chapter 3. Methodology

3.1 CRISP-DM:

The study was carried out using the methodology CRISP-DM (Cross-Industry Standard Data Mining Process). This methodology offers a study framework that helps the findings to be better and quicker. The CRISP-DM methodology organizes the study into six stages, these stages assist to better comprehend the process and provide a road map for planning and conducting the research. The following figure shows the CRISP-DM model phases.

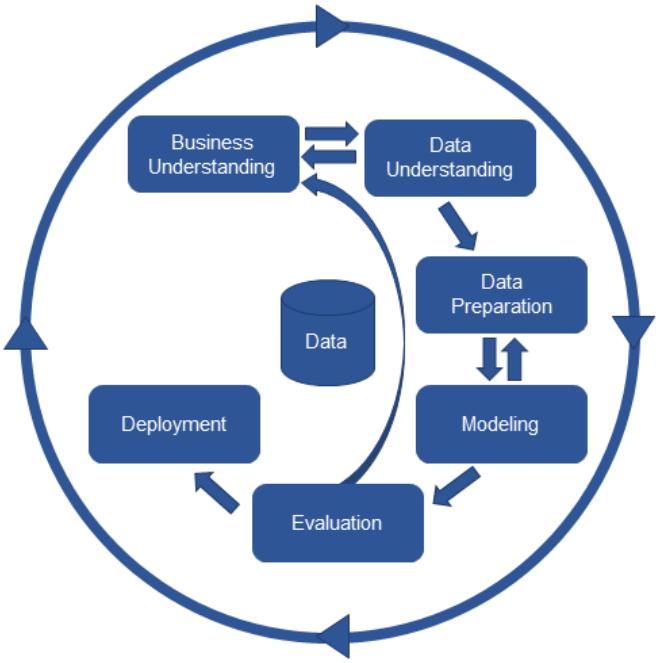


Fig. 3.1 CRISP-DM Methodology

The arrows show the flow of the process and the frequent dependencies between the phases. The phases of the CRISP-DM methodology are as follows:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

Each of the phases of the CRISP-DM model will be approached in detail with respect to the research.

3.2 Business Understanding

The first and most significant stages of the research project is the business understanding, this stage is aimed at understanding the goals of the project from a business viewpoint and turning this business view into a definition of study problems and then creating a plan for achieving these goals.

This phase essentially implies what we are attempting to resolve and what is the client need or issue that we are attempting to explain, and afterward relying upon the sort of the issue that we are attempting to comprehend, at that point we can perceive what sort of arrangement we can use to take care of the issue. This phase involves finding the type of problem, in this research project the problem is to analyse the news which are coming out on Retail Industry. After that we can make a proper step to figure out the issue.

when we have distinguished the issue that we need to resolve by means of Machine learning. It must be referenced here that not all issues should be settled by means of ML, henceforth we have to observe cautiously what the business needs is here in this circumstance and we have to represent the entirety of the suspicions, objectives just as requirements also, that this business issue accompanies so we can locate an answer for this.

3.3 Data Understanding

In this process the acquisition of data is more important to build efficient machine learning algorithms. In this phase we must perceive what sort of information that we have. In this project data is collected from Webhose API: https://webhose.io/. Webhose uses specific querying technique to fetch JSON data. The data is in large blocks of json files.

There are numerous measurable properties that we can watch that may add the synopsis measurements too, although there are values that are clear or nulls and how would we have to deal with these, and furthermore if any anomalies and how would we have to deal with these. We can likewise perceive how all the parameters are associated with one another to observe its importance to remaining parameters and to check whether there a solid connection between certain parameters.

3.4 Data Preparation

The data preparation phase consists of all the activities required to process the data from the initial raw form before it is fed into the model. These tasks include everything from attribute selection to transforming and cleaning the data. Report Data contains unstructured data. We cannot therefore include the raw test data as an input to the classification.

The data preparation is one of the stages in the methodology process and is also the important stage where the acquired data is transformed into pure quality for the sake of building better machine learning models. The extracted data needs to be pre-processed to understand by machine learning algorithms for classifying the data into specific category with better accuracy. The data is stored in the json format.

The data stored in json format is converted into CSV file with python programming language and then concatenated all the CSV file to make final CSV. Furthermore, the converted data is then processed with several steps such as checking the null values, missing values, changing the categorical variables, adding the attributes to make a final dataset.

3.4.1 Pre processing using NLTK

The first step is to prepare input data for the models, which involves removing stop words, punctuations, and unnecessary numbers. For the purpose of this experiment, we decided to use Natural Language Tool Kit (NLTK) from SKLearn library to help with the process. Word embedding is the collective name for a set of language modelling and feature Techniques for learning in natural language processing (NLP) where vocabulary terms or phrases are mapped to vectors of real numbers.

Initially, we need to tokenize the document into words to operate on word level. Text data will be noisy. So, we need to drop the number of words. In addition, text data may contain numbers, unwanted white spaces, tabs, punctuation characters, stop words etc. We also need to clean data by removing the unwanted. The transformed data is then contrasted with the words in a predefined dictionary, which consisted of particular words and phrases in terms of feelings and their related polarity power. To avoid words that occur in only one or two documents, we use a minimum frequency of documents that takes into account words that appear in at least three documents. Null values in news column was replaced by the news of the previous day.

Some sample entries are as below

clean_text	original_text
bayside shoes life diabetic shoes shoes people...	Bayside Shoes for Life\nDiabetic Shoes and sho...
covid impacts reshapes retail last months expe...	COVID-19 impacts & reshapes retail Over the la...
retail sales arena entry level sales position ...	Retail Sales Job Arena\nEntry level sales posi...
managing director nati harpaz exits catch nati..	Managing director Nati Harpaz exits Catch\nNat...
house home stages home environs everything sal...	House to Home stages home environs where ever...

Table 3.4.1 News Before and after cleaning

3.5 Modelling

In the modelling phase, several models selected based on thorough research are applied. Each of models has a specific requirement on how the data must be pre-processed, hence the need to step back into the data processing phase. This phase includes the selection, creation and assessments of the models.

In this research, the models are initially implemented on the training set list to forecast for the next day These forecasted values are compared with the test set list values and the test accuracy of the models are computed. Once the test errors are computed, the entire data set list is fed into the models for the forecast. The best models for forecasting are selected based on the research conducted in this field.

3.6 Deployment

In the last phase we deploy the model. It includes all the necessary code modules for the project to execute in a single package which is readily available and can be easily accessed by other developers. The developed code modules for this project is published in the cloud environment or server environment by stacking the code modules into a compressed packaged which is generally referred to as deploying the code in the release mode. The code modules for this project is developed using the python platform and coded in python programming language. The code modules published in the cloud environment is easily accessible and hence it can be deployed within the local system to run the code with an ease.

Chapter 4. System Design and Specifications

In this chapter, we will be elaborating the system configuration and specification details that we have used for developing the machine learning model.

Operating System: Windows 10 Operating System

Processor: Core i5

Programming Language: We have used python programming for developing the application. The python language is predominantly used for development of project in extracting the data, performing pre-processing on the extracted data, building the machine learning algorithms.

Python Packages: The names of the python programming package and the libraries used in the above-mentioned process are provided here.

Pickle Library: In most of the cases, the trained model weights need to serialize and stored in a file or any other format in a system, so that the trained model can be used for performing the prediction without training the model instantly for every predictions. To do this, we need to serialize the model weight, which can be perform using the functions in the pickle library. It converts the data into character stream and the same can be deserialized into python object using the scripts **(Fasnacht, 2018).**

Math Library: The python Math library provides access to the common math functions that can be used to perform complex mathematical calculations and access to the constants in python. We do not have to explicitly install this library, as it is included by default in the python module **(Bergstra and Breuleux, 2020).**

Pandas: Pandas is a Python package that is used to perform analysis and manipulation on the data. It holds the data like in the excel file and it is called as Data Frames. The pandas rely on some of the other packages like Numpy and MatPlotLib **(McKinney, n.d.).**

Numpy: Numpy is a python package that is used to perform the scientific calculations. The multi-dimensional container of data can be maintained with the help of numpy package. Lists is a package that is like numpy and can be used as an alternative **(McKinney, n.d.).**

 # Plotly: Plotly.py is a Python-based collaborative, open-source, and browser based graphing library. Plotly.py is a high-level, declarative charting module, built on top of plotly.js. Plotly.js ships with more than 30 forms of charts including science charts, 3D charts, statistical charts, SVG maps, financial charts and more.

 # Keras: Keras is an open-source neural networking library written in Python. Atop it can run TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or PlaidML.

 # Seaborn: Seaborn is a matplotlib-based Python data visualisation library. It provides a high-level interface for drawing attractive statistical graphics and providing information.

 # Spacy: SpaCy is an open source software library, written in Python programming languages, for advanced nlp.

 # Nltk: NLTK is a main gathering for creating Python projects to work with information from the human language. It offers simple to-utilize interfaces for more than 50 corporate and lexical devices, for example, WordNet, alongside a set-up of text-preparing libraries for characterization, tokenisation, following, labeling , parsing, and semantic thinking.

 # Sklearn: Scikit-learn (originally scikits.learn, and also known as sklearn) is a free machine learning library for Python language programming software.

 # wordcloud: A Word cloud (or Tag cloud) is a snapshot of text data. It shows a list of words, with the size or colour of each font being displayed. This format is useful to get the most popular words understood quickly.

#matplotlib: Matplotlib is a simple library for visualization that allows man to run and plot other libraries at its base like seaborn or word cloud.

Chapter 5. Implementation

In the implementation chapter, we will be discussing in detail on the machine learning models that we have used in our research work for the implementation. On top of that, we will be explaining about the user interface of the application that we have developed using the Python flask web framework. We have also mentioned the list of the packages that we have used in the Python programming and the purpose of using the same in our application.

5.1 MACHINE LEARNING MODEL:

The machine learning programming is used for developing a model that can understand the pattern among the data and make the decision with respect to the user input. In our case, we extract the data from one of news data collection website: https://webhose.io/. Webhose uses specific querying technique to fetch JSON data. To implement the discussed research, we have taken the consideration of these machine learning models which are listed below.

- K-means
- LDA

The machine learning models has been chosen on based of the literature review papers that we have analysed in the literature review chapter. We know the fact that no machine learning model is best and the type of data, the volume of the data and the other related factors decides which machine learning algorithm best fits for our case. So, it is essential to take multiple machine learning models for a single research work implementation and proceed with doing the enhancements and deployment with the one that has shown the best accuracy and performance.

5.2 K-MEANS ALGORITHM

K-MEANS proposed by Geon is one of the most common, partitioning based clustering methods. K-means' disadvantages are that it needs one to initially set the quantity of groups and afterward pick the underlying bunching focuses. Idle Dirichlet Allocation (LDA) is an adult probabilistic subject model that assists with decreasing dimensionality, semantine mining and recovery of data in text. We present a LDA and K-implies (LDA K-implies)-based grouping approach for papers. To help the bunching impact of reports with K-implies, we find the underlying grouping communities by making sense of the standard inert points separated by LDA. The viability of the LDA K-implies is surveyed on the informational collections of the 20 Newsgroups. We show that LDA K-implies, rather than grouping[10]

dependent on arbitrary instatement of K-means and LDA, can altogether improve the bunching impact.

The K-implies calculation is one of the apportioned based bunching calculations that was prevalently utilized in regions, for example, data recovery and customized suggestion. In any case, starting bunching focuses are haphazardly picked in the customary K-implies calculation. The bunch results are consequently excessively reliant on the underlying grouping places, especially when records are spoken to with a pack of - words model (BOW). Archives are frequently spoken to as high-dimensional and scanty vectors when utilizing crude terms as highlights – two or three thousand measurements and a sparsity of 95 to 99 percent is typical[8].

In such cases, in the event that we use K-implies in report grouping, we should tackle two issues: One is the means by which to lessen the dimensionality of the data set and catch more framework semi direct as productively as could reasonably be expected; another is the way to find the underlying bunching focuses that can speak to most idle bunch semantic data.

5.2.1 Network Model of K-Means

1. Each node with an ID is numbered

2. They are all fixed or pseudo-static nodes.

3. Both nodes will send out the data to the BS.

4. Their energy consumption is regulated by all nodes.

5. The initial energy is the same for all the nodes.

6. The CHs are mindful of their resources left over.

7. The sensor nodes are distributed at random within the target region.

5.2.2 Steps of K-Means

Step 1: Initial clustering

K-means algorithm is performed with aim WSN for cluster creation. Suppose the n nodes WSN is divided into clusters k. Next, it randomly selects k out of n nodes as the CHs.

According to the Euclidean distance each of the remaining nodes determines the CH closest to it.

Step 2: Re-clustering

The centroid of each cluster is determined after each of the nodes in the network is allocated to one of the clusters k. Step 2 is recursively implemented with the new CH in each cluster, until the CH is no longer modified.

Step 3: Choosing the CH

If the clusters have been created, each node of a cluster is allocated an ID number according to the distance from the centroid, assigning a smaller number to the closer one. A node's ID number shows the order to be selected as CH. Thus, the ID number plays a significant role in selecting a node as CH.

5.2.3 ELBOW METHOD

Elbow method is a technique that looks at the percentage of variance as a function of the number of clusters described. This approach is based on the premise that a number of clusters should be chosen so that adding another cluster does not offer any better data modelling. The percentage of variance which the clusters describe is plotted against the number of clusters. Most information will be added by the first clusters but at some stage the marginal gain will drop dramatically and give an angle in the graph. At this point the right "k" i.e. number of clusters is chosen, thus the "elbow criterion". [13]The idea is that start with K=2, and proceed to increase it by 1 in each phase, calculating your clusters and the cost of the training. At some value for K the cost drops significantly, and when you increase it further it reaches a plateau afterwards. This is your ideal K-value. The reasoning is that you are increasing the number of clusters after this because the new cluster is very similar to some of the current ones. In that fig. 1 The distortion J) (descends rapidly with K increasing from 1 to 2, and from 2 to 3, and then W hits the elbow at K=3, and then the distortion goes down very slowly. And then it looks like the correct number of clusters is maybe using three clusters, since that's the elbow of this curve. Distortion goes down rapidly until K=3, and actually goes down very slowly after the number of clusters needed for this collection of data is 3.

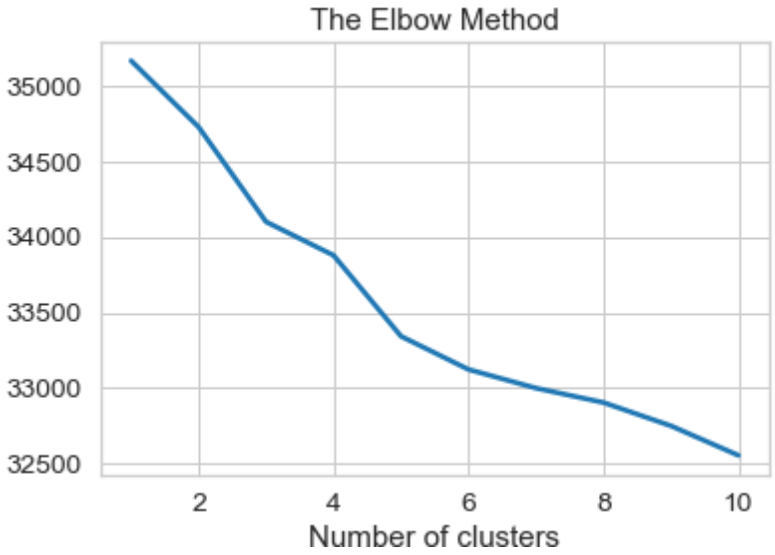


Fig 5.2.3 Elbow Method

Algorithm

1. Elbow Approach for defining K-means 1. Let k=1 initialize

2. Starting

3. Increase the value of k

4. Measure the expense of high quality solution

5. When the cost of the solution drops significantly at some point

6. That is the real k.

7. Finish

The cluster nodes start the computations based on pre-evaluated cluster number and divide themselves in the clusters according to the pre-evaluation. The cluster nodes divide themselves by Euclidean distance calculation into the pre-evaluated number of clusters. The creation of the clusters is carried out using the K-Means algorithm[12]. K-Means algorithm has already been proven faster in cluster nodes than the LEACH algorithm. LEACH algorithm continues to be the most balanced and efficient algorithm to most WSN scenarios. Based on the location coordinates of each node (X & Y) the K-means use the Euclidean distance in two ways field. The K-means then determines the distance formula and divides the cluster nodes into the number of clusters in particular.

5.3 Latent Dirichlet Allocation for Topic Modelling

A text such as – Word Frequency and Inverse Document Frequency – has several approaches to obtain topics. Techniques for factorizing Non-Negative Matrix. Latent Dirichlet Allocation is the most common modelling technique for topics and we'll address the same in this paper.

LDA believes that records are derived from a variety of subjects. Those subjects then produce words based on their distribution of probabilities[12]. LDA backtracks and tries to find out what issues will generate such documents in the first place, given a dataset of documents.

LDA is a method of factorisation of matrixes. Any corpus (document collection) can be interpreted as a document-terminal matrix in vector space. The following matrix displays a corpus of N documents D1, D2, D3 ... Dn and M word size W1,W2 .. Wn. Wn. The value of I j cell in Document Di gives the frequency count of word Wj.

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

Table 5.3 corpus of N Documents

This Document-Term Matrix is transformed by LDA into two lower dimensional matrices – M1 and M2. M1 is a matrix of document-topics and M2 is a subject – a matrix of words with dimensions (N, K) and (K, M) respectively, where N is the number of documents, K is the number of topics and M is the size of the expression.

	K1	K2	K3	k
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

Table 5.3.1 Document-Term Matrix

Notice that these two matrices already have topic distributions for word and document topics, but this distribution needs to be enhanced, which is LDA's main objective. To boost these matrices LDA makes use of sampling techniques.

It Iterates for each document "d" for every word "w" and attempts to change the current subject – word assignment with a new assignment. A new subject "k" is assigned to word "w" with a probability P that is a combination of two probabilities p1 and p2.

Two probabilities are calculated for each topic: p1 and p2.

P1 – p(topic t / document d) = the proportion of words currently assigned to topic t in document d.

P2 – p(word w / topic t) = the proportion of subject t assignments to all documents that originate from that word w.

The current topic – word assignment with probability, combination of p1 and p2 is replaced with a new subject. The model assumes that at this point, all existing word-topic assignments, except the current one, are correct. This is basically the likelihood that word w was created by subject t, so it makes sense to change the topic of the current word with new likelihood.

A steady state is reached after a number of iterations, where the subject of the document and the subject term distributions are reasonably strong. This is the point of LDA convergence.

Parameters of LDA

Alpha and Beta hyperparameters - alpha represents the density of the document-topic and Beta represents the density of the subject-word. Higher the alpha value, more topics are composed of documents and lower the alpha value, less subjects are found in documents. On the other hand, the higher the beta, the topics are made up of a large number of terms in the corpus, and with the lower beta value, they are made up of few terms[14].

Number of topics-Number of topics from the corpus to be collected. By using the Kullback Leibler Divergence Rate, researchers developed approaches for obtaining an optimal number of topics.

Topic Terms Number-Number of words composed in a single subject. It's usually determined by the criteria. If the problem statement talks about extracting themes or ideas, it is recommended that a larger number be chosen if a problem statement talks about extracting features or phrases, it is recommended that a lower number.

Number of iterations / passes – Maximum number of iterations required to converge to the LDA algorithm.

Latent: This refers to everything we don't know a priori, and is hidden in the details. The themes or topics that make up the document are unknown here, but they are assumed to be present as the text is produced based on those topics.

Dirichlet: It's a 'distribution of distributions.' Yes, we read that correctly. But what that means? Let's think about an example on this. Suppose there is a machine making dice, and we can monitor whether the machine will always produce a dice of equal weight on both sides, or whether there will be bias on certain sides. So the dice-producing system is a distribution as it generates dice of various kinds. We also know that the dice itself is a distribution, so when we roll a dice we get several values. This is what being a set of distributions means and that is what Dirichlet is. Here the Dirichlet is the distribution of topics in documents and the distribution of terms in the subject in the sense of topic modelling.

Allocation: This means that once we have Dirichlet we can assign topics to the topics of the document's documents and terms.

That is what it is. That is, in a nutshell, what LDA is. Now let's understand how this works in modelling of topics.

The Algorithm

LDA is a type of unsupervised learning , it considers documents as bags of words (i.e., no matter how order is). LDA makes a key assumption: picking a set of topics and then picking a set of words for each topic was the way a document was produced[15]. Now we could ask, "Well, how do we find the topics

"Well, the answer is simple: engineers are reversing that mechanism For each document m it does the following for this purpose:

1. Suppose there are k subjects in all the records.
2. Distribute these k points through report m (this dissemination is alluded to as α and might be symmetric or uneven, more on this later) by doling out a theme to each term.
3. For each word w in document m, assume its topic is incorrect but the correct topic is assigned to every other word.
4. Probabilistically relegate word w to a two-dimensional theme:

    - What are the subjects in report m
    - How ordinarily word w has been relegated to a particular subject in all the reports (this dissemination is called β, more on this later).

5. Repeat this step for each document many times, and we are finished!

The Model

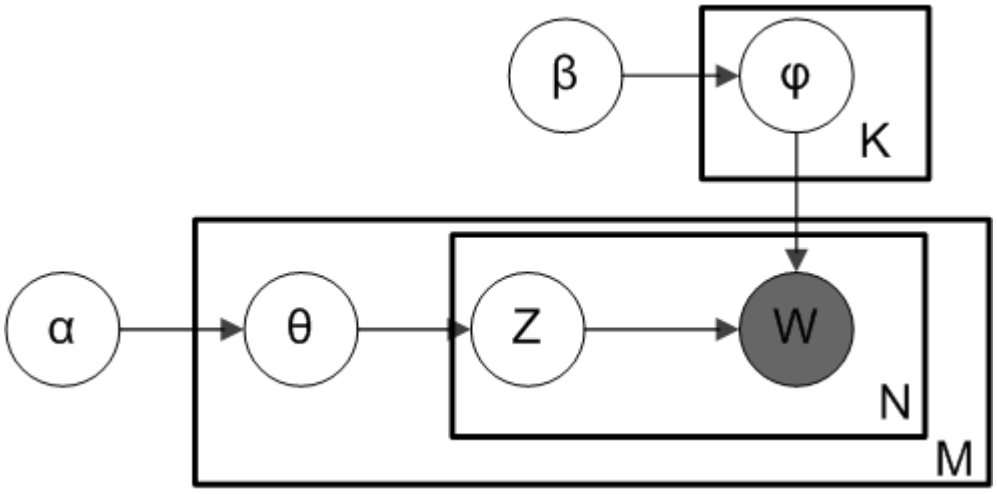


Fig 5.3 LDA Model

Above is what is known as an LDA model plate diagram, where:

α - Is distributions of topics per document,

β - A per-topic distribution of words,

θ - is the topic of document m distribution,

φ - is the word distribution for subject k,

z - is the topic of the n-th term in document m, and

w - is a specific word

5.3.1 LDAvis: A visualizing method and topic interpreting

We present LDAvis, an interactive web-based visualization of topics calculated using the Latent Dirichlet Allocation, which is constructed using a combination of R and D3[2][3]. Our visualization offers a global view of the topics (and how they vary from each other), while simultaneously allowing for a detailed analysis of the words most closely associated with each subject.

First, we propose a novel method for selecting the words to present to a consumer in order to assist in the topic interpretation process, in which we identify the relevance of a term to a subject. Secondly, we present the results of a user study which suggests that ranking terms purely by their likelihood under a topic is suboptimal for topic interpretation. We implement an interactive visualization framework that we call LDAvis, which attempts to address some basic questions about a topic model that is fitted:
Latent Dirichlet Allocation (LDA) (Gardner et al . , 2010; Chaney and Blei, 2012; Chuang et al., 2012b; Gretarsson et al . , 2011) has been given a lot of attention recently to visualizing the performance of the topic models fit. Due to the high dimensionality of the designed model, such visualizations are difficult to construct – LDA is usually applied to several thousands of documents which are modelled as mixtures of dozens (or hundreds) of topics, which are themselves modelled as distributions over thousands of terms (Blei et al., 2003; Griffiths and Steyvers, 2004). Interactivity is the most promising basic technique for producing compact and detailed LDA visualisations.
We implement an interactive visualization framework that we call LDAvis, which attempts to address some basic questions about a topic model that is fitted:
(1) What does every topic mean?
(2) How prevalent is any topic? And,
(3) How do the various topics relate?

Each of these questions is answered by different visual elements, some of which are original and some of which are borrowed from existing resources.

There are two fundamental elements of our visualisation. Firstly, our visualization 's left panel shows a global view of the subject model, and addresses questions 2 and 3. In this perspective, we plot the topics as circles in the two-dimensional plane whose centres are calculated by measuring the distance between subjects, and then using multi-dimensional scaling to translate the inter-topic distances into two dimensions, as is done in (Chuang et al., 2012a). Using the areas of the circles, we encode the overall prevalence of each subject, where we group the topics in decreasing order.

Second, the right panel of our visualization depicts a horizontal bar chart whose bars reflect the individual words most useful for understanding the topic currently selected on the left, and allows users to address question 1, "What is the meaning of each subject? ". A pair of overlaid bars represent both a given term's corpus-wide frequency and the term's topic-specific frequency, as in (Chuang et al., 2012b).

Our visualization's left and right panels are connected in such a way that selecting a subject (on the left) shows the most useful words (on the right) for understanding the chosen topic. Furthermore, the selection of a term (on the right) shows the conditional distribution for the chosen term over topics (on the left). This kind of linked collection enables users to look in a compact way at a large number of topic-term relationships.

A key innovation of our system is how we deter the most useful words for understanding a given subject, and how we allow users to change that determination interactively. A subject in LDA is a multi-nomial distribution of words in the corpus' vocabulary over (typically thousands of). Using anywhere from three to thirty terms in the vocabulary, one usually uses a ranked list of the most possible words in that field to describe a subject. The problem with viewing topics this way is that common words in the corpus frequently appear for several topics near the top of such lists, making it difficult to distinguish the definitions of these topics.

5.3.2 Relevance of terms to topics

Definition of Relevance

The most common way of presenting a topic, a discrete distribution of words, is to print out the top ten words ordered within this subject by decreasing frequency. There is nothing much more we can do given a single subject. But knowing other subjects which describe the same

corpus gives us more knowledge. We seem to be able to use this knowledge to select appropriate words to describe topics.

Word distinctiveness and saliency were designed not for a particular subject, but to find appropriate words corpus-wide. Finding applicants for subject representation isn't good for them. We present a word relevance score in this section within a framework based on the same idea: penalize the word frequency by a metric that measures how much the word is exchanged across topics.

First, we consider the frequency p(w) of the word within a topic k, instead of the global word frequency p(w|k). Then as a penalty for sharing, we divide by the exponential entropy e^{H_w} where

$$H_w \triangleq -\sum_k p(k \mid w) \log p(k \mid w) \qquad \textbf{Eq 1}$$

is the entropy of the distribution of subjects given a word w, representing the degree to which the word w is spread across many subject areas. We identify measure of relevance

$$\mathcal{R}(w \mid k) \triangleq \frac{p(w|k)}{e^{H_w}} \qquad \textbf{Eq 2}$$

for word w as the frequency divided by the exponential entropy within subject k.

5.3.3 The LDAvis System

An online, intuitive representation framework, LDAvis, that takes into consideration a careful examination of the point term connections in a LDA model, while at the same time giving a worldwide perspective on the subjects in a conservative space through their pervasiveness and similitudes. We additionally recommend a novel measure, significance, to rank terms inside themes to help with the subject understanding crucial, we present outcomes from a client study demonstrating that positioning terms in diminishing request of probability are problematic for point translation.

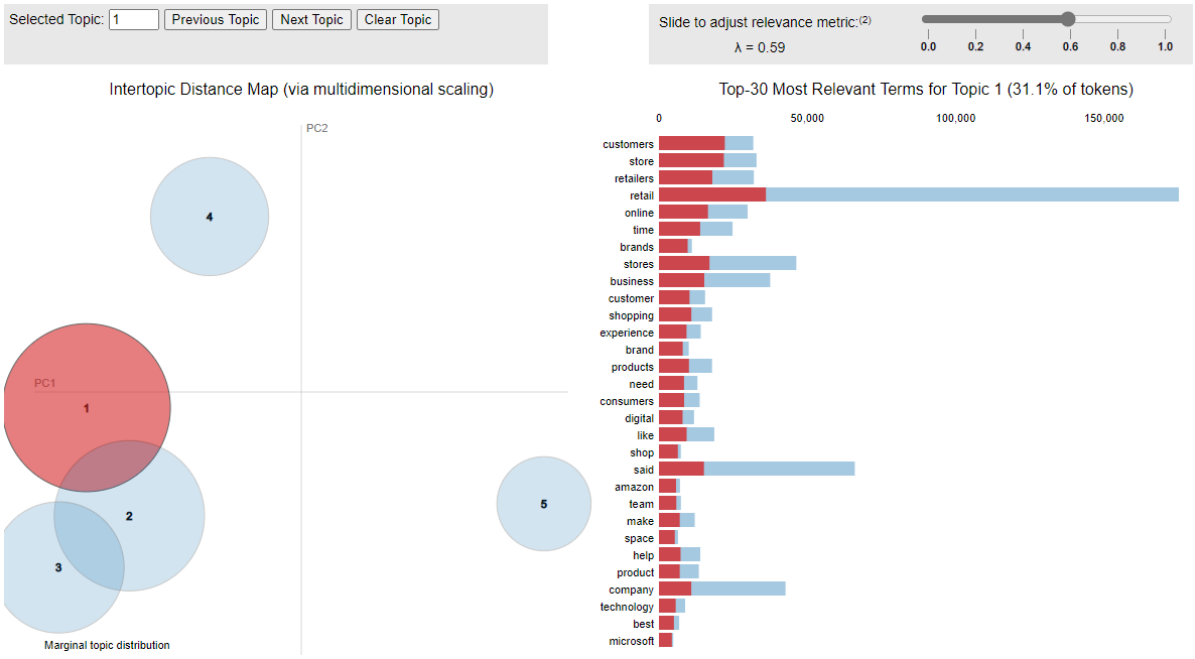


Fig 5.3.3 LDA visualisation

For this topic, when we set that relevance factor, it finds the importance of words irrespective of occurrence. If we decrease the relevancy and go for a topic the keywords will change. So for the topic we can set the relevancy level, when we set the relevancy level the proportion of the occurrence changes.

So if we select topic 4, we are setting the relevance factor to 1 and $\lambda = 0.27$, it says the topic depends majorly on market, so topic 2 and topic 5 spoke about market but they are not market centric, that's where this principal components coming in to the picture(pc1 & pc2).

We can have multiple principal components as well in any directions, Ideally when you are concentrating at this principal component (pc1), we can say that all these are fallen in to this pc1. If we can see perpendicularly, this topic 4 and topic 5 are totally out of box.

If we go for topic 5 with relevance factor $\lambda = 1$, we will know what words are majorly occurring. It is talking about retail, shares, company. If we highlight retail, it will be equal in all the topics, almost 3 is concentrating on something else like health, covid.

Labelling - All these belongs to one category of words, that is what it segregates, that is topic 3. Now we will see topic 1 major number of documents 31.1% of the whole documents and topic 2 24%.

Chapter 6. Testing and Evaluation

In this chapter, we will be discussing on the test executions with the machine learning model that we have build and the results we have obtained at the end of the execution. On top of that, we will be discussing on the evaluation phase that we have undergone for the implemented model.

6.1 Project implementation steps

1. Clean and conduct on data an Exploratory Data Analysis (EDA).
2. Vectorisation of text data (Count Vectorizer and TF-IDF) cleaned.
3. Generate a Word Cloud to see what are the most commonly spoken words.
4. Preform Topic Modelling to see if we can find any simple, different topics on which people speak.
5. Use clustering methods to cluster trends from text data to see if we can cluster.

6.2 Design Diagram:

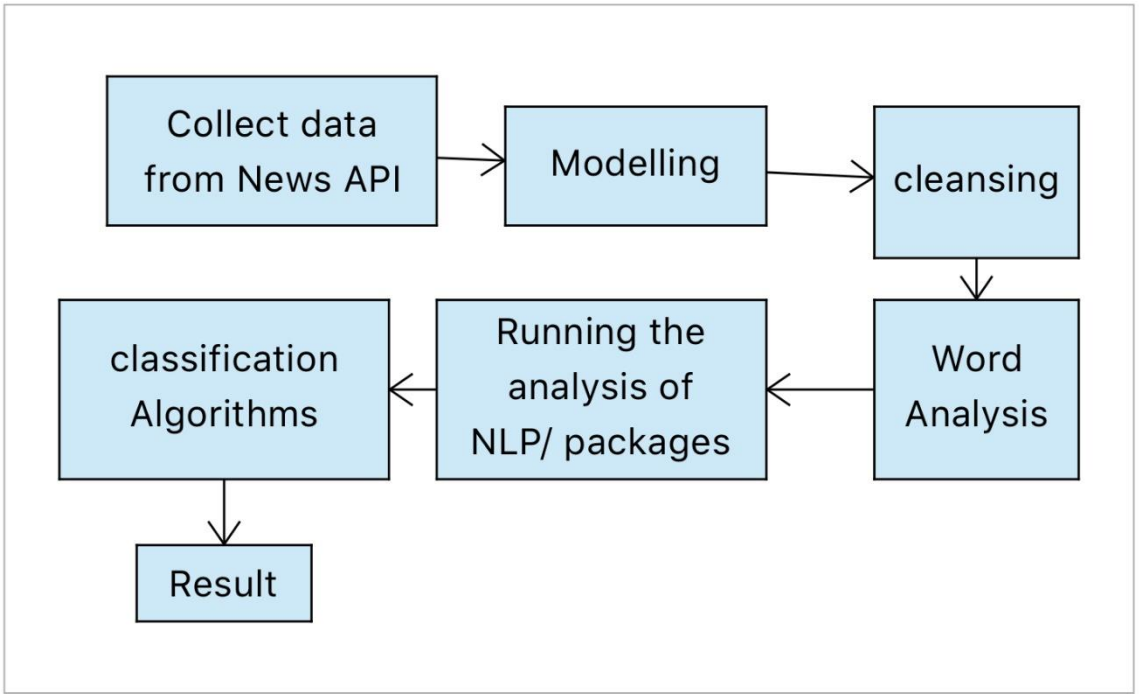


Fig 6.2 Design Diagram

6.3 Data extraction from Webhose API

We extracted the data from one of news data collection website: https://webhose.io/. We got 4 months of news for retail related keywords as seen below. Webhose uses specific querying technique to fetch JSON data.

Keywords:

- Customer Resource Management
- Integrated Supply Chain
- Supply chain management
- Visual merchandising
- merchandising
- Warehouse management system
- Cross Merchandising
- Consignment Merchandise
- Destination Retailer
- Franchise
- Franchis
- Franchisee
- High-Speed Retail
- Inventory Management
- Market Penetration
- Omni-Channel Retail
- Social Commerce
- Tribetailing
- niche retailing
- Inventory Management
- Inventory Turnover

Query used to fetch data from webhose:

title:"retail" -text:('Customer Resource Management' OR 'Integrated Supply Chain' OR 'Supply chain management' OR 'Visual merchandising' OR 'merchandising' OR 'Warehouse management system' OR 'Cross Merchandising' OR 'Consignment Merchandise' OR

'Destination Retailer' OR 'Franchise'OR 'Franchis' OR 'Franchisee' OR 'High-Speed Retail' OR 'Inventory Management' OR 'Market Penetration' OR 'Market Research' OR 'Omni-Channel Retail' OR 'Social Commerce' OR 'Tribetailing' OR 'niche retailing' OR 'Inventory Management' OR 'Inventory Turnover') language:english site_type:news

6.4 Data Transformation

The Data transformatin phase involves flattening the data from JSON files for each month to a single .CSV file

Input files:

- 17747_webhose_2020_04_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json
- 17747_webhose_2020_05_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json
- 17747_webhose_2020_06_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json
- 17747_webhose_2020_06_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json

Output file:

- mergedjsondata.csv

6.5 Data Pre-processing

We'll take the following steps:

- Tokenisation: break the text into phrases and phrases into words; Lower the words and avoid punctuation.
- Delete words with less than 3 letters.
- They delete all the stop words.
- Words are lemmatized — third-person words are changed to first-person words, and verbs are converted in past and future tenses into present.

- Words are truncated — words are reduced to the root form.

	uuid	clean_text	original_text
0	0bec306c531c33455f6b41296f27f53c96c5f5c6	bayside shoes life diabetic shoes shoes people...	Bayside Shoes for Life\nDiabetic Shoes and sho...
1	977e7f90f8f26a3d8261b72ab097e39f47f2b329	covid impacts reshapes retail last months expe...	COVID-19 impacts & reshapes retail Over the la...
2	1e397b4b493bfe5084bf2a09b2792305251fd1da	retail sales arena entry level sales position ...	Retail Sales Job Arena\nEntry level sales posi...
3	c46b5d01a08d66c21ad5f4a3f09ddeb508797841	managing director nati harpaz exits catch nati...	Managing director Nati Harpaz exits Catch\nNat...
4	a7f8fa953c23a5c6eb20f5232d429e0adf863e75	house home stages home environs everything sal...	House to Home stages home environs where ever...

Fig 6.5 Data Pre-processing

6.6 K-Means

In clustering k-means, each observation — each text for our purposes — can be allocated to one, and only one, cluster. But the subject models are the models of mixtures. That means that each document is assigned a likelihood of belonging to a latent theme or "topic."

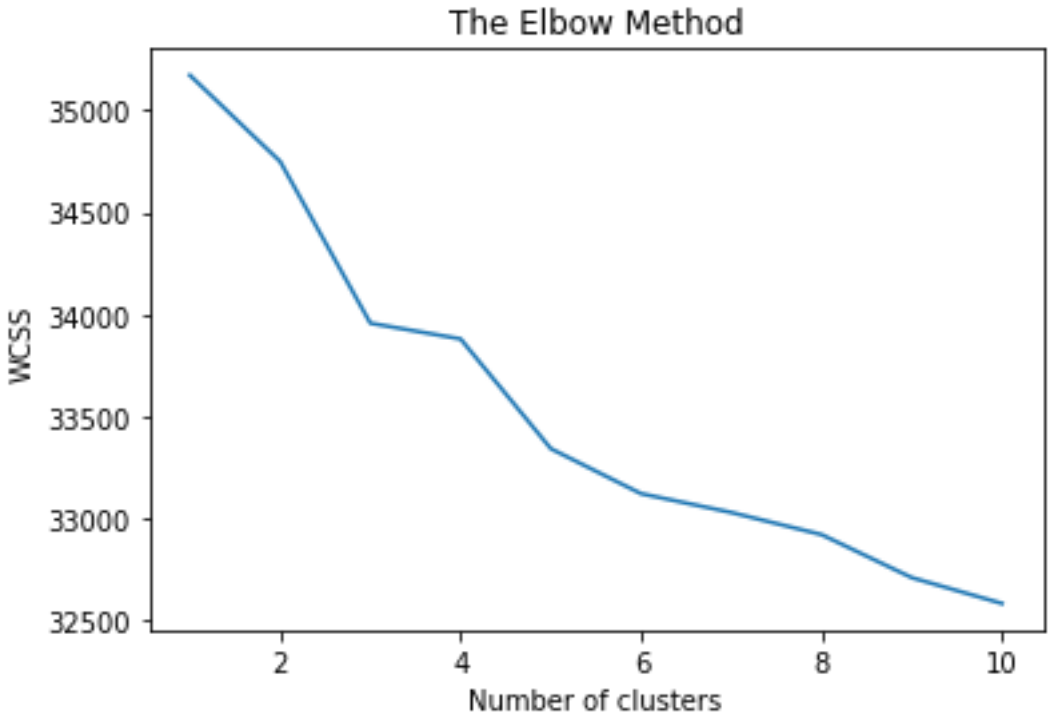


Fig 6.6 K- means Elbow curve

For a number of values for k (say from 1-10) the k-means clustering will be run by elbow method on the dataset and then for each value of k calculates an average score for all clusters. The distortion score is determined by default, the number of square distances between each point and its assigned centre.

So as to choose the ideal number of groups, we should pick the estimation of k at the "elbow" for example the point after which the mutilation/inactivity starts to diminish straightly. In this

manner we reason that the ideal number of bunches for the information is 3 for the given information.

6.7 LDA Implementation

The entire code can be found as a Jupyter Notebook.

1. Data Loading
2. Cleaning up data
3. Exploratory Analysis
4. Preparing LDA-analysis data
5. Training on LDA model
6. Analysing the effects of the LDA model.

6.8 Exploratory Data Analysis

Exploratory data analysis (EDA) is a systematic way to explore the data using transformation and visualization. EDA is an iterative cycle and it's not a process with any set of rules however some basic steps to be applied that help to manage data in a systematic way:

Step 1: Generate questions about data.

Step 2: Search for answers by visualizing, transforming, and modelling the data.

Step 3: Use what is to learn to refine questions and if required then generate new questions.

To verify that the pre-processing has happened properly, we can build a word cloud using the word cloud kit to obtain a visual representation of the most common words. It is important to understand the data and to ensure that we are on the right track, and whether further pre-processing is required before the model is educated.

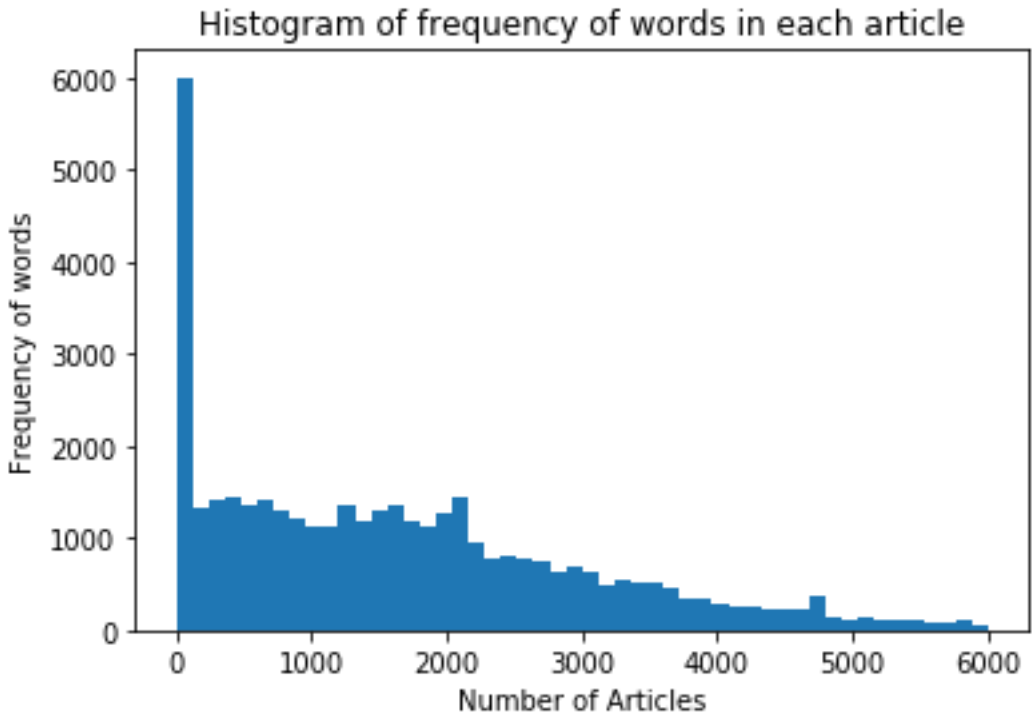


Fig 6.8 Histogram of frequency of words in each article

Sentiment analysis is perhaps one of NLP's most common technologies, with a large array of tutorials, courses, and technologies centered on analyzing the feelings of different datasets ranging from corporate surveys to film reviews. The main element of analyzing sentiment is evaluating a body of text to understand the perspective it expresses. Usually, this emotion is quantified with a positive or negative meaning, called polarity. Overall opinion is also derived from the polarity score sign as positive, neutral , or negative.

Typically, an interpretation of emotions works better on text with a subjective meaning than on text with an objective context only. Objective text usually portrays any normal statements or facts without voicing any thoughts , emotions or moods. Subjective text includes text normally conveyed by a person who has normal moods, thoughts , and feelings. Analysis of sentiments is commonly used, particularly as part of social media analysis for any domain, whether it is a company, a recent film, or a product launch, to understand people's reception and what they think of it based on their opinions or, it, feelings!

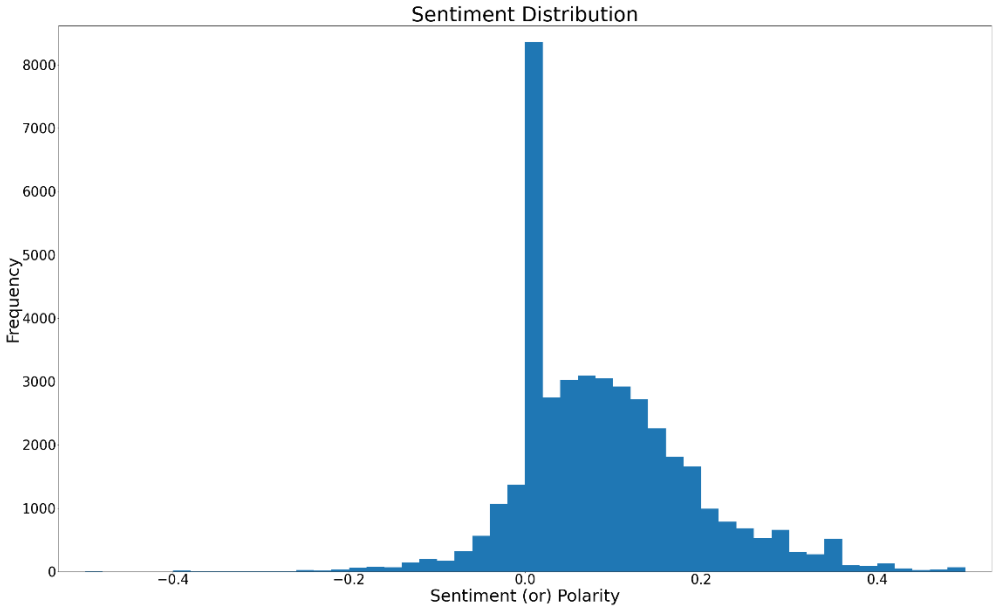


Figure 6.8.1 Sentiment Distribution

We may have seen several times a cloud filled with lots of words in different sizes, reflecting the frequency or value of each word. This is known as Tag Cloud or WordCloud. You can learn how to build your own WordCloud in Python for this tutorial, and customize it as you see fit. This tool is going to be very useful to explore the text data and make your report more vibrant.

Fig 6.8.2 Word Cloud

6.9 Prepare text for LDA Analysis

Next, let 's center around making an interpretation of printed information into an arrangement that will fill in as a contribution to the LDA model for preparing. We start by changing over

the reports to a basic portrayal of vectors (Bag of Words BOW). To begin with, we'll transform a rundown of titles into vector records, all of which have lengths equivalent to the jargon. At that point we will plot the ten most generally utilized terms dependent on the consequence of this methodology (the content vector list). In the word cloud these words ought to likewise happen as a check.

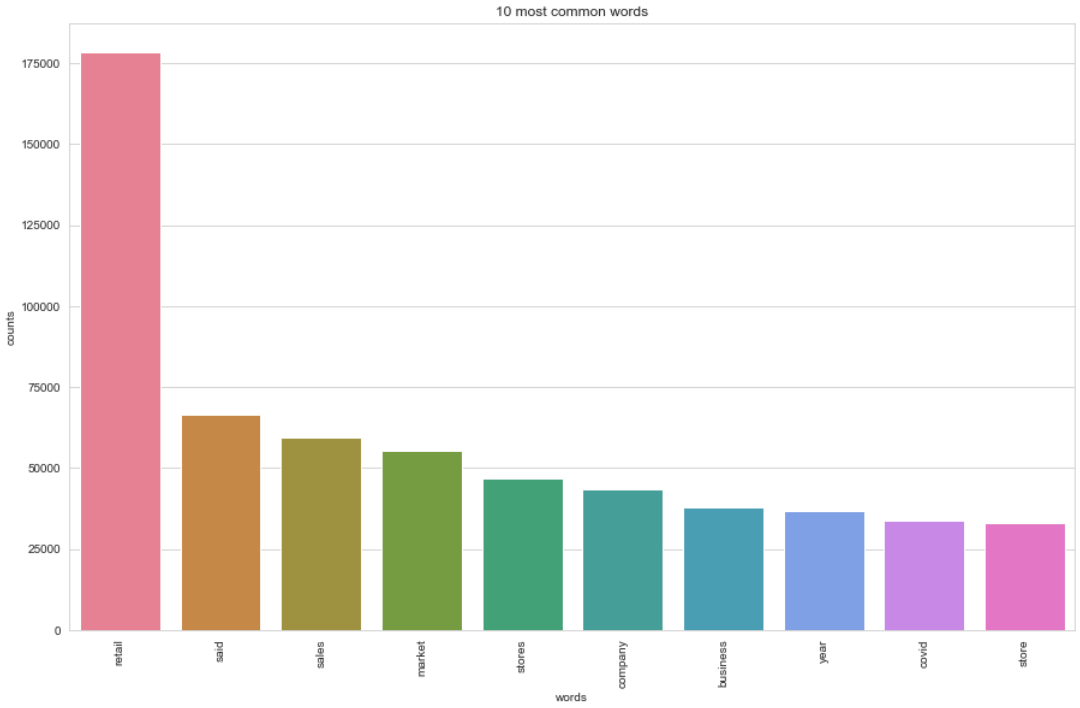
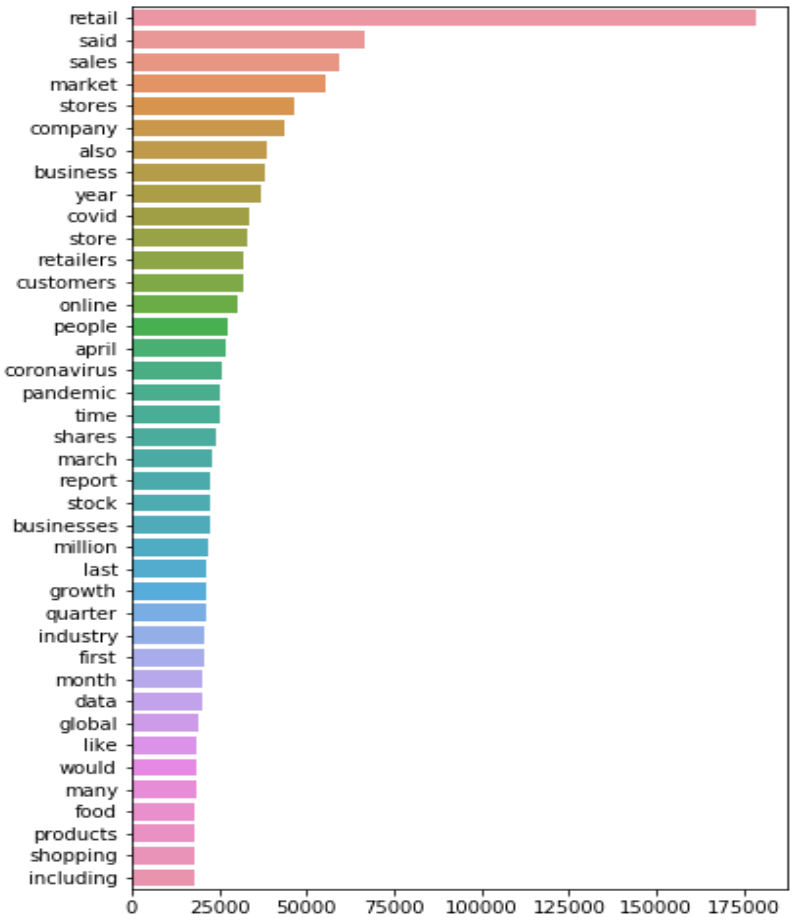


Fig 6.9 10 Most Common Words

Bar charts are useful to display counts, or to summarize statistics with error bars. Matplotlib provides quality-publishing statistics across platforms in a range of hardcopy formats and virtual environments. Python scripts, Python and IPython terminal, web application servers, and various graphical user interface toolkits can be used with Matplotlib.



6.9.1 Matplotlib Bar chart

6.10 LDA model training and visualization results

Now we will see the model training and the topics which are found by LDA.

```
Topics found via LDA:

Topic #0:
said retail stores covid businesses people health coronavirus state open

Topic #1:
retail store customers experience online customer time retailers people love

Topic #2:
retail market global growth business table report industry company product

Topic #3:
retail shares company stock properties quarter rating price investment research

Topic #4:
retail sales year said stores month april percent march cent
```

Fig 6.10 Topics found via LDA

6.11 LDA model results

Since we have a prepared model let 's envision the interpretability subjects. To do that, we 're going to utilize a typical representation unit, pyLDAvis, intended to intuitively help with:

1. Better understanding and perception of topic, and
2. Clear interpretation of the thematic relationships.

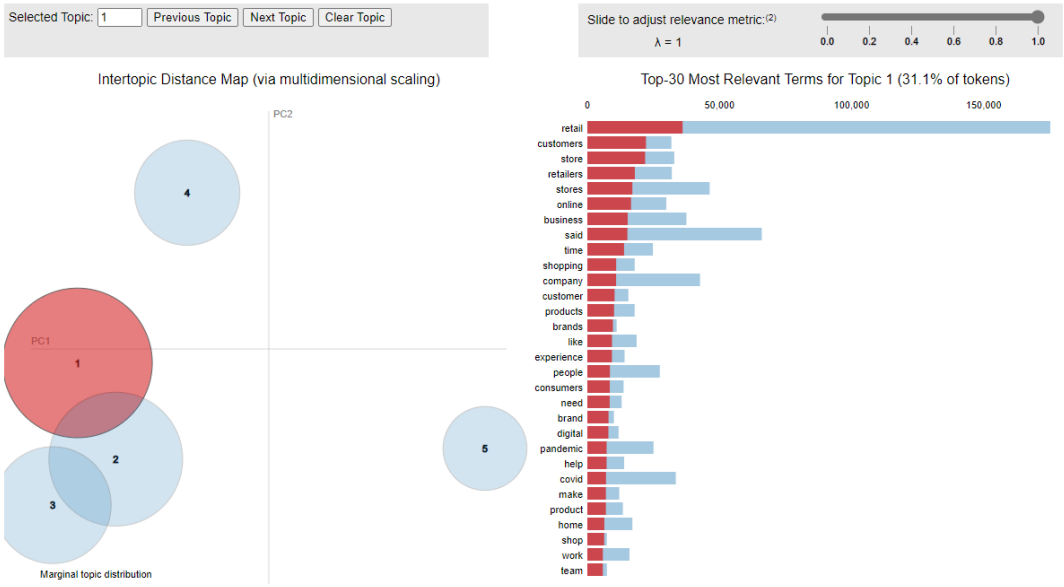


Fig. 6.11 LDA model results

For (1), you can pick every subject manually to see its generally normal or potentially "related" top terms utilizing diverse boundary esteems. This can help when you attempt to give each subject a human interpretable name or "setting."

For (2), investigating the Intertopic Distance Plot will assist you with seeing how themes identify with one another, including the conceivable structure of more significant level subject classes.

Chapter 7. Conclusions and Future Work

Over the past decade, machine learning has become increasingly popular, and ongoing improvements in computational accessibility have prompted exponential development for individuals looking for approaches to execute new techniques to propel the normal language preparing field.

We began with understanding what modelling of the topic can do. Using LDA, we developed a simple theme model and visualized the topics using pyLDAvis. Then we developed LDA implementation from mallet. We have shown how to find the optimum number of topics using coherence scores and how you can arrive at a reasonable understanding of how to pick the optimum model.

Finally we saw how the findings could be aggregated and summarized to create ideas that could be in a more actionable way.

In Future, we foresee completing a more extensive client examination to more readily see how to advance point understanding in fitted LDA models, including a correlation of different strategies, for example, the Turbo Topics positioning (Blei and Lafferty, 2009) or the FREX scores (Bischof and Airoldi,, notwithstanding significance. We additionally note the need to envision associations between subjects, as this will give understanding into what's going on at the degree of the archive without really introducing entire reports

References

1. MEDIUM. 2020. **Topic Modelling : Art of Storytelling in NLP** - Medium. [online] Available at: < https://medium.com/@MageshDominator/topic-modeling-art-of-storytelling-in-nlp-4dc83e96a987/>.
2. 2020.[online] Available at: <https://www.researchgate.net/publication/303563965_A_Text_Mining_Research_Based_on_LDA_Topic_Modelling > .
3. Jason Chuang, Christopher D. Manning, Jeffrey Heer Stanford University Computer Science Department{jcchuang, manning, jheer}@cs.stanford.edu. Visualization Techniques for Assessing Textual Topic Models 2018 IEEE International Conference on Big Data (Big Data).
4. Carson Sievert, Iowa State University and Kenneth E. Shirley AT&T Labs Research. LDAvis: A method for visualizing and interpreting topics
5. Loulwah AlSumait, Daniel Barbara, James Gentle, and Carlotta Domeniconi. 2009. Topic Significance Ranking of LDA Generative Models. ECML.
6. Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture [online] Available at: <https://www.researchgate.net/publication/324482968 2020 >.
7. AYTUĞ ONAN1, SERDAR KORUKOĞLU2, AND HASAN BULUT. 2016. LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis, International Journal of Computational Linguistics and Applications vol. 7, no. 1, 2016, pp. 101–119.
8. Rubayyi Alghamdi and Khalid Alfalqi. *(IJACSA).*A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications, Vol. 6, No. 1, 2015.*
9. Peng Guan, Yuefen Wang, Bikun Chen, Zhu Fu, School of Economics & Management, Nanjing University of Science & Technology. K-means Document Clustering Based on Latent Dirichlet Allocation Available at<https://www.semanticscholar.org/paper/K-means-Document-Clustering-Based-on-Latent-Guan/9e623a64d1d3f8f73bfedc855c3b8f6861eea591>

10. Chonghui Guo, Menglin Lu & Wei Wei. 2019, An Improved LDA Topic Modeling Method Based on Partition for Medium and Long Texts. Available at<https://link.springer.com/article/10.1007/s40745-019-00218-3#citeas>
11. Hamed Jelodar , Yongli Wang , Chi Yuan , Xia Feng, Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Available at<https://www.ccs.neu.edu/home/vip/teach/DMcourse/5_topicmodel_summ/notes_slides/LDA_survey_1711.04305.pdf>
12. AnalyticsVidhya.2020, Beginners Guide to Topic Modeling in Python. Available at <https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/>
13. TowardsDataScience 2020 Topic Modeling in Python: Latent Dirichlet Allocation (LDA). Available at <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>
14. Susan Li. 2018. Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. Available athttps://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
15. Towards Data Science. Tyler Doll. 2018. LDA Topic Modeling: An Explanation. Available athttps://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd
16. Ayoub Bagheri, Mohamad Saraee, Franciska de Jong. ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. June 11, 2014; Issue published: October 1, 2014. Available at<https://doi.org/10.1177/0165551514538744>
17. G. Xu, Y. Zhang and X. Yi. Modelling User Behaviour for Web Recommendation Using LDA Model. *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, NSW, 2008, pp. 529-532, doi: 10.1109/WIIAT.2008.313.
18. Sergey I. Nikolenko, Sergei Koltcov, Olessia Koltsova. Topic modelling for qualitative studies. December 2015. Available at <https://doi.org/10.1177/0165551515617393>
19. Yaswanth Kalepalli , Pasupuleti Durga Phani Teja, Suneetha Manne . 2020. Effective Comparison of LDA with LSA for Topic Modelling. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). Available at < https://ieeexplore.ieee.org/abstract/document/9120888>.

20. Muzafar Rasool, Bhat,Majid A Kundroo,Tanveer A Tarray &Basant Agarwal .2019. Deep LDA : A new way to topic model. Available at<https://doi.org/10.1080/02522667.2019.1616911>
21. Shaymaa H. Mohammed , Salam Al-augby. 2019. LSA & LDA Topic Modeling Classification: Comparison study on E-books. Indonesian Journal of Electrical Engineering and Computer Science Vol. 19 , No. 1, Jul 2020.
22. M.S. Saranya ; P Geetha. 2020. Word Cloud Generation on Clothing Reviews using Topic Model. International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0177-0180, doi: 10.1109/ICCSP48568.2020.9182111.
23. Nidhi Singh, Nonita Sharma, Ajay K. Sharma, Akanksha and Juneja. 2018. Sentiment Score Analysis and Topic Modelling for GST Implementation in India. Available at <https://doi.org/10.1007/978-981-13-1595-4_19>.

Appendix I

This document will guide you through the contents of the Artifacts and the necessary steps to implement the python code for dissertation project titled “Analysing Retail Sector News using NLP during COVID-19”.

Contents of the Artifacts

Query used to fetch data from webhose:

title:"retail" -text:('Customer Resource Management' OR 'Integrated Supply Chain' OR 'Supply chain management' OR 'Visual merchandising' OR 'merchandising' OR 'Warehouse management system' OR 'Cross Merchandising' OR 'Consignment Merchandise' OR 'Destination Retailer' OR 'Franchise'OR 'Franchis' OR 'Franchisee' OR 'High-Speed Retail' OR 'Inventory Management' OR 'Market Penetration' OR 'Market Research' OR 'Omni-Channel Retail' OR 'Social Commerce' OR 'Tribetailing' OR 'niche retailing' OR 'Inventory Management' OR 'Inventory Turnover') language:english site_type:news

Input files:

- 17747_webhose_2020_04_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json
- 17747_webhose_2020_05_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json
- 17747_webhose_2020_06_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json
- 17747_webhose_2020_06_9f0f66b0d8a5cf37ba19f9ffa9db57b3_0000001.json

Output file:

- mergedjsondata.csv

Python Code:

- The entire code can be found as a Jupyter Notebook.

Readme

- Explains the contents of the Artefacts, how to implement the code on python.