Topic Modeling on News Articles

Team members

Ajinkya Morade

Nitesh Gajakosh

Abstract

In the Web era, people tend to rely on the Web to receive news instead of traditional ways such as newspapers. However, the amount of news generated online is enormous that prohibits people from obtaining their interested news. Most of the common newswire sites still classified the news manually that costed a lot of human effort and may receive unstable result. In the past decades, text classification has been a hot topic and received attention from many scholars in areas such as natural language processing, information retrieval, and machine learning, etc. Various classification algorithms and models have been developed to tackle this problem. In the meantime, Tim Berners-Lee proposed the concept of linked data in 2006.

Key words

Machine learning, Text classification, News classification

1.Introduction

Online news services have emerged in past decades due to the pervasion of mobile devices and Internet access. Many readers, especially youngsters, relied on online news services rather than traditional sources such as newspapers, TV, and radio. A survey in 2017 reported that 41 percent of Internet users from the United States named TV as their main source of news, whereas 44 percent stated the internet (incl. social media) was their main news source.1 Similar shares were also applied across countries worldwide. Such emergence of online news makes it difficult for users to receive their interested news among a sea of news generated constantly. Therefore, newswires gave class labels or tags to categorize the news articles to allow users to access their interested news easily. However, such tasks were tedious and subjective, and may produce questionable result which may diminish the user experience and satisfaction. Research on automatic news classification was thus attract attention from researchers from various areas. The aim of automatic news classification is to assign a predefined class to each news article according to its similarity to the class representative measured in some metric. A major source for such similarity measurement is the content of the news articles, which is mostly textual. Therefore, techniques of text classification or text mining were commonly adopted to tackle the news classification tasks.

Text classification methodologies rely on the comprehension of the semantics behind the texts to produce precise classification. In generally, the semantics of a piece of text is realized as the set of important keywords occurred in this piece of text. Several approaches on the identification and assessing of the keywords have been suggested, e.g. the vector space model

2.Problem Description

In this project your task is to identify major themes/topics across a collection of BBC news article. You can use clustering algorithms such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) etc.

2.1 Data Description

The dataset contains a set of news articles for each major segment consisting of business, entertainment, politics, sports and technology. You need to create an aggregate dataset of all the news articles and perform topic modeling on this dataset. Verify whether these topics correspond to the different tags available.

3.EDA on given Data set

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

3.1 Data Analysis:

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

3.2 Data Sourcing

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

3.3 Data Pre-processing

A dataset may contain noise, missing values, and inconsistent data; thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

3.4 Data Cleaning

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

Missing Values

1. Incorrect Format
2. Incorrect Headers
3. Anomalies/Outliers

3.5 Data Deduplication

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

3.6 Missing Values

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of service and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.
2. MAR (Missing at random): These values may be dependent on some other features.
3. MNAR (Missing not at random): These missing values have some reason for why they are missing.

3.7 Dropping Missing Values

One of the ways to handle missing values is to simply remove them from our dataset. We have known that we can use the is null () and not null () functions from the panda's library to determine null values.

3.8 Measures of Central Tendency

The measure of central tendency tends to describe the average or mean value of datasets that is supposed to provide an optimal summarization of the entire set of measurements. This value is a number that is in some way central to the set. The most common measures for analysing the distribution frequency of data are the mean, median, and mode.

3.9 Measures of Dispersion

The second type of descriptive statistics is the measure of dispersion, also known as a measure of variability. If we are analysing the dataset closely, sometimes, the mean/average might not be the best representation of the data because it will vary when there are large variations between the data. In such a case, a measure of dispersion will represent the variability in a dataset much more accurately.

Multiple techniques provide the measures of dispersion in our dataset. Some commonly used methods are standard deviation (or variance), the minimum and maximum values of the variables, range, kurtosis, and skewness.

3.10 Standardizing Values:

To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale. For example, if the data contains the values of the top speed of different companies' cars, then the whole column should be either in meters/sec scale or miles/sec scale.

3.11 Univariate analysis

Univariate analysis is the simplest form of analysing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

3.12 Bivariate Analysis

If we analyse data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

a) Numeric-Numeric Analysis:

Analysing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyse it in three different ways.

- Scatter Plot
- Pair Plot
- Correlation Matrix

b) Numeric - Categorical Analysis:

Analysing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyse those mainly using mean, median, and box plots.

3.16 Multivariate Analysis

Multivariate analysis is the analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behaviour more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

3.17 Correlation Among Variables

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour

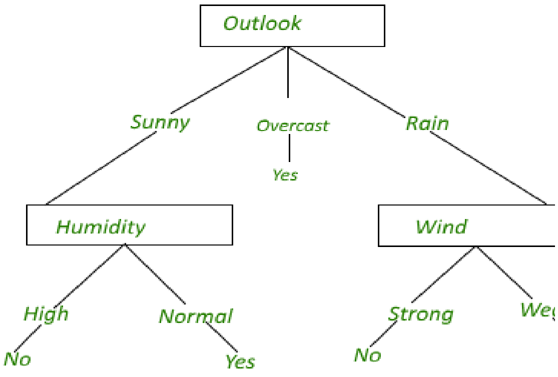
3.18 Graphical Representation of The Results

This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. This is also an essential step as the result analysed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA. Most of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

4.Algorithms

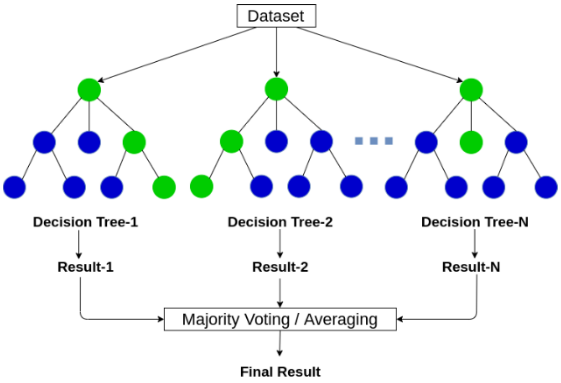
4.1. Decision Tree:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called *recursive* partitioning. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.



4.2. Random Forest:

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the greatest number of times a label has been predicted out of all.

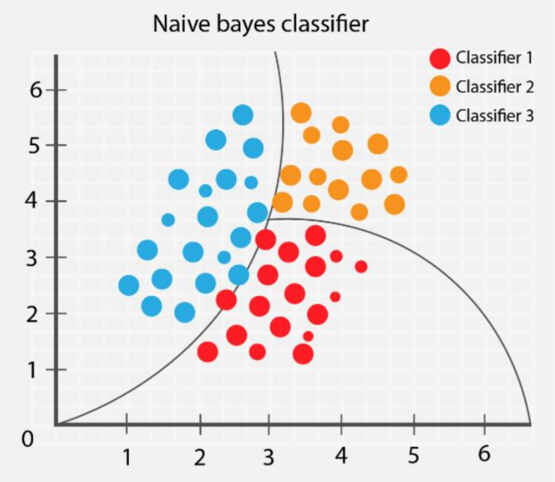


4.3. Navie Bayes:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

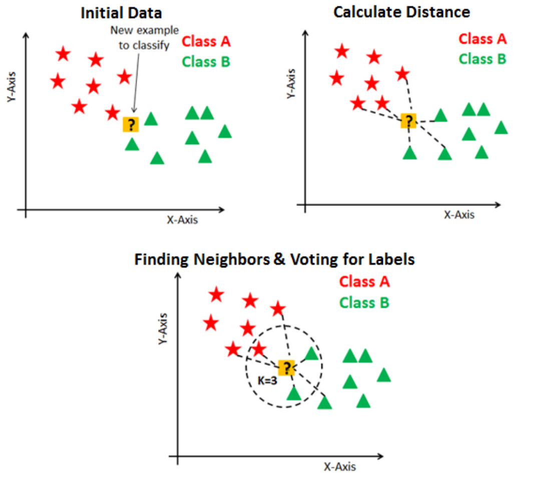
For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.



4.4.KNN Classifier:

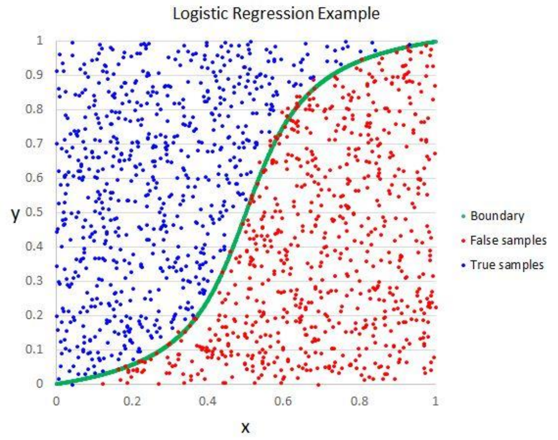
K Nearest Neighbour (KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won’t vote. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.



4.5 Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). In statistics the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value)



5.Conclusions

After implementing various models on the given data such as Logistic Regression, Decision Tree Classifier, Naïve Bays Classifier, KNN Classifier, Random Forest Classifier. We get maximum accuracy with Logistic Regression and KNN Classifier train and test accuracy are 0.98, 0.95 and 096, 0.95 for Logistic Regression and KNN Classifier respectively. Both the models are of different types as Logistic Regression works on math base where KNN is a distance based it looks for its nearest neighbours.

6.References

[1] “Content Enrichment Using Linked Open Data for News Classification” by Hsin-Chang Yang, Yu-Chih Wang.

[2]” Classification of News Articles using Supervised Machine Learning Approach” by Muhammad Imran Asad, Muhammad

Abubakar siddique, Safdar Hussain, Hafiz Naveed Hassan, and Jam Munawwar Gul.

[3]” Text classification of BBC news articles and text summarization using text rank” by Abhishek Dutt and Kirk Smalley.