UNIVERSITETI I EVROPËS JUGLINDORE
УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА
SOUTH EAST EUROPEAN UNIVERSITY

FAKULTETI I SHKENCAVE DHE TEKNOLOGJIVE BASHKËKOHORE
ФАКУЛТЕТ ЗА СОВРЕМЕНИ НАУКИ И ТЕХНОЛОГИИ
FACULTY OF CONTEMPORARY SCIENCES AND TECHNOLOGIES

POSTGRADUATE STUDIES-SECOND CYCLE

THESIS:

"Data Curation System for Fake-News Dataset Generation"

CANDIDATE:
Seadin Osmani

MENTOR:
Assoc. Prof. Dr. Visar Shehu

Tetovo, 2021

UNIVERSITETI I EVROPËS JUGLINDORE
УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА
SOUTH EAST EUROPEAN UNIVERSITY

FAKULTETI I SHKENCAVE DHE TEKNOLOGJIVE BASHKËKOHORE
ФАКУЛТЕТ ЗА СОВРЕМЕНИ НАУКИ И ТЕХНОЛОГИИ.
FACULTY OF CONTEMPORARY SCIENCES AND TECHNOLOGIES

POSTGRADUATE STUDIES-SECOND CYCLE

THESIS:

“Data Curation System for Fake-News Dataset Generation”

CANDIDATE:
Seadin Osmani

MENTOR:
Assoc. Prof. Dr. Visar Shehu

Tetovo, 2021

Personal Statement

I, Seadin Osmani, Student in the Software Engineering study program, offered by the Faculty of Contemporary Sciences and Technologies for the second cycle of studies, state that the Master Thesis titled “Data Curation System for Fake-News Dataset Generation” is written by me, and it is original and unique. The sources used are cited on a regular basis.

Acknowledgment

Sincere thanks to my mentor, professor Visar Shehu, and to the commission members, professors Adrian Besimi and Besnik Selimi for the help, support, and encouragement throughout the process of writing this Master Thesis. Also, special thanks to my family for always being there for me.

Abstract

Fake News is a phrase used to refer to misleading or incorrect information, namely information that is fabricated and not verified by facts or reliable sources.

Fake News or misinformation in recent years has become a very serious issue, especially with the massive expansion of social media platforms and the information-sharing ability these platforms provide to everyone.

This issue has seen an influential impact in sensitive cases such as Political Elections around the world, disinformation about the development of new technologies, and most importantly, the surge of fake news recently about the Global Pandemic of Covid-19 that has impacted the entire world.

This work will provide an analytics-driven approach for gathering news data from various real and fake sources, storing the data as datasets, processing the data, and concluding with a result.

Keywords: fake news, data, information, dataset

Abstrakt

Lajme të rreme si frazë përdoret për të përshkruar informacione të pasakta që janë të fabrikuara, të pa verifikuara dhe të pambështetura në fakte ose burime të besueshme.

Lajmet e rreme ose dezinformimi në vitet e fundit është bërë një shqetësim shumë serioz, veçanërisht me përhapjen masive të platformave të mediave sociale dhe ofrimin e mundësise nga këto platforma për shpërndarje të informacionve, që ato e mundësojnë për secilin.

Ky problem ka pasur ndikim të theksuar në raste të ndjeshme siç janë zgjedhjet politike në mbarë botën, dezinformimi në lidhje me zhvillimin e teknologjive të reja, dhe më e rëndësishmja, vërshimi i lajmeve të rreme në lidhje me Pandeminë Globale të Covid-19 që ka ndikuar të gjithë botën.

Ky punim do të ofrojë një qasje analitike për mbledhjen e të lajmeve nga burime të ndryshme reale dhe të rreme, ruajtjen e te dhenave si datasete, përpunimin e të dhënave, dhe konkludimin me një rezultat.

Fjalë kyçe: lajme të rreme, të dhëna, informacione, dataset

Апстракт

Лажни вести е израз кој се користи за да се опишат неточни или лажни информации што се измислени, непроверени или неподдржани од веродостојни факти или извори.

Лажните вести или дезинформациите во последниве години станаа многу сериозен проблем, особено со масовното ширење на платформите на социјалните мрежи и можноста за размена на информации, што овие платформи го овозможуваат за секого.

Овој проблем имал значително влијание врз чувствителни случаи како што се Политички избори низ целиот свет, дезинформации за развојот на нови технологии и што е најважно, поплавата на лажни вести за Глобалната пандемија на Ковид 19, која влијаеше на целиот свет.

Ова теза ќе обезбеди аналитички пристап кон собирањето податоци од различни извори на вести, чување како датасети, обработка на податоците и заклучок со еден резултат.

Клучни зборови: лажни вести, податоци, информации, датасет

Contents

Table of Figures

1. Introduction

Misinformation, otherwise described as "Fake News" in recent years, has become a Global problem, putting people's approach to news in a difficult situation.

Although many definitions have been given to further describe this issue, the short and clear definition would be that Fake News is information that lacks truthfulness that is published and shared to an audience, or even worse, it is fabricated propaganda with bad intentions towards a person, a community, or a movement.

While rumors, gossip, and satire are not new forms of journalism, they never reached the level of concern that Fake News did with its impact and influence in many important and sensitive topics like politics and health these past years.

Although considerable progress has been made in the battle against fake news, it is still significantly present as a phenomenon.

The important challenges that it has brought to society in present times include the challenge to make individuals more aware of what they read and consume online, which might affect the way they perceive the world, the fight to completely stop and prevent misinformation from being distributed in the future and raising the trust in professional journalism.

In this Thesis, we will try to analyze and get an overview of news in the Albanian language in present times by building a tool that will gather news articles from various real and fake news sources. With the help of algorithms, we will process the data and will try to achieve an analytical result for both types of news.

2. Background Research and State of the Art

2.1. Background Research

Fake News is a relatively new expression that has become popular in recent years. Although its use in recent times has changed meaning and form, the term is currently used both as a disrespectful term denouncing media and journalism, and as a term for various forms of false, misguided, or fabricated information leading to unintentional misinformation and or intentional propaganda, according to [1].

The answer to why fake news has become such a big problem in recent years lies in the massive rise of Social Media platforms. According to [2], 1.3 million new users joined these online platforms during the year 2020 alone, meaning 15 new users every second. A total of 4.20 billion people are now users of social media around the world, which to the Digital Media are a massive audience and easy to reach Target.

According to [3], the incentive to create and fabricate misinformation is mainly backed by two factors, to fulfill a social agenda, like provoking a certain community against a political candidate or cause, or to earn revenue online through Advertising by getting visitors to their sites, mainly by misleading the users with eye-catching and clickbait titles.

As reported by a study performed by the Pew Research Center, In the USA, this issue (fake news) is ranked there as a larger problem than other even more worrying issues like racism, climate change, or terrorism [4]

In the year 2020, a global pandemic hit the entire world, causing panic and insecurity about what was about to happen. This was a great opportunity for fake news publishers. Hot topics such as the safeness of the Vaccine, the health situation concerns linked to the new 5G telecommunication technology, and many other topics influenced social media, causing a Journalism Pandemic in itself.

With a problematic issue like this, an analytics-driven approach will be the focus of work by gathering data from several sources and processing the data by using various tools and algorithms.

2.2. State of the Art

The attempts and efforts to fight and stop fake news have started since late 2016, when they had a huge impact on the Donald Trump vs. Hillary Clinton race.

Social Media, although a great way of getting informed, was heavily abused by this phenomenon up to the point that concerns were raised about what is real and what is not, served to the audience about many sensitive situations.

Although independent fact-checking projects and initiatives acted, it soon was realized that they were not enough. Thus Facebook, as one of the most impacted platforms by fake news, took other measures and steps to fight them. Facebook took measures by modifying its News Feed algorithm many times as a crucial way of spreading fake news by preventing them in the first place. [5]

Because a lot of fake news is financially motivated, spammers cover their content with ads. To fight these kinds of fake news, Facebook acted with the help of Machine learning to assist their response teams by updating automatic detection and enforced its policies against inauthentic spam accounts. It also took measures to make it hard for people posting fake news to buy ads on its platform. Another measure was modifying their News Feed raking algorithm and making it easier for users to report false stories. [6]

Twitter, another big social media platform impacted by massive waves of misinformation, took measures too. A special project was initiated by Twitter lately called “Birdwatch” to try and fight misinformation by adding an option for people to identify information in tweets they believe are inaccurate and write informative contextual notes. [7]

3. Motivation, Problem Statement, and Research Goal

3.1. Motivation

The 2016 US elections were a turning point for the spread of fake news on Social Media sites like Facebook and Twitter, having caused serious concerns about trust in institutions and a feeling of threatened democracy. The concern is based on the ability of content being relayed among users and communities without any fact-checking or filtering, according to [8]

According to a study by [9], in the final months of that election, the most popular fake news stories on Facebook received more engagement and viewership than stories from relevant big news corporations. This, in principle, may have led to an impact on the final election results.

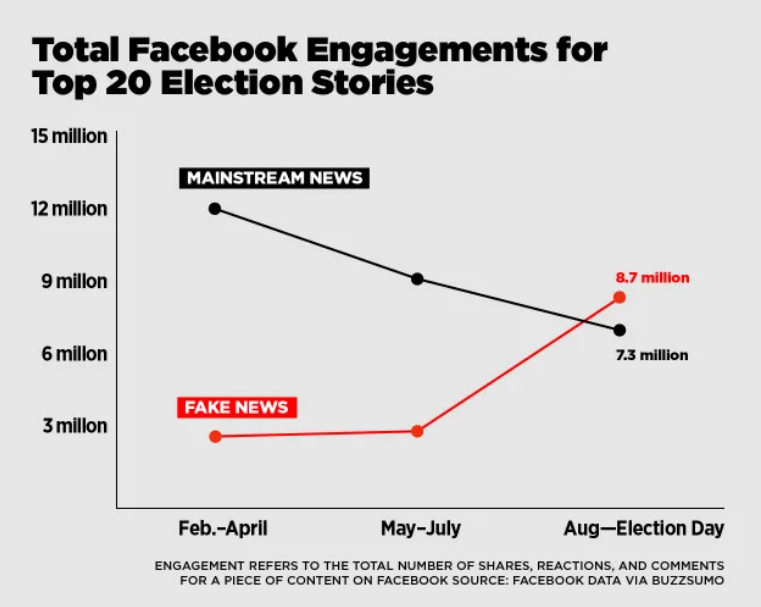

Figure 1. Analysis for Total Engagement of Top 20 Election Stories

Source: [9]

A study by [10] finds that many of those viral fake news stories had their epicenter of distribution, surprisingly in the small city of Veles, in our country North Macedonia.

The people behind this activity were youngsters who reportedly made millions in revenue during the pre-election period, just by posting viral fake news on social media, targeted to a more conservative audience in the USA.

3.2. Problem Statement

The global pandemic that was a shock to the world in 2020, affecting the lives of everyone and causing people to remain isolated to prevent the spread of the deadly virus that caused many victims, brought the problem of Fake News even more to attention. This was a perfect time for News Distributors online as the confusion and panic were at a high level of uncertainty.

As the pandemic spread, social media platforms emerged as an important form of socializing during the uncertainty. The usage of social media increased up to 87% around the world during that period. A study shows that in the Country of Italy, which was one of the most impacted at the beginning of 2020, 46.000 news posts were shared daily that were either inaccurate or linked to some form of misinformation about the crisis.

Examples of such false stories included topics such as "5G" technology having caused Covid-19, the transmission of the virus by getting bit by certain insects, and unverified absurd natural cures to such a deadly virus, and many other conspiracy theories. All these stories were fake.

After analyzing 1225 fake stories, social media was responsible for about half of them (50.5%). The other half consisted of sources like individuals, websites, newspapers, and surprisingly important politicians. [11]

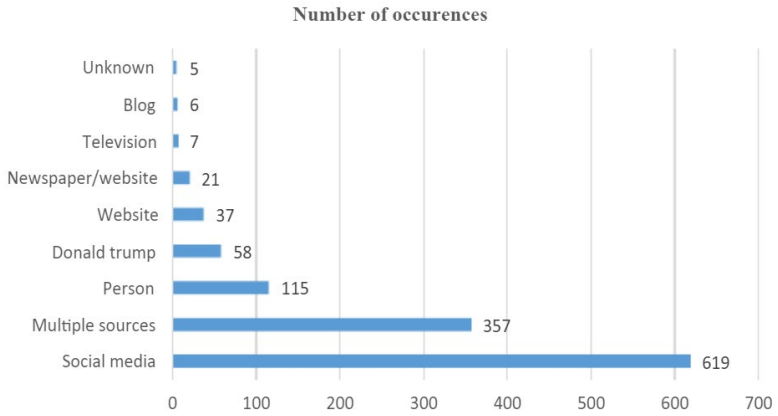

Figure 2. Sources of fake news about the pandemic

Source: [11]

According to the same study, another finding was about the time when the False news reached their peak. It was exactly the time people were more uncertain, right when the pandemic lockdown took place, in the middle of March.

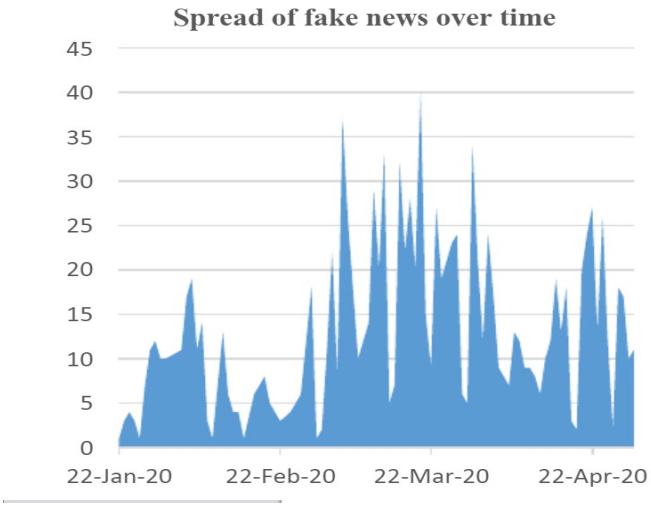

Figure 3. Spread of fake stories between months January and April 2020

Source: [11]

All the above mentioned led the World Health Organization, UN, and other important Organizations to warn about an ongoing "infodemic" with lots of misinformation during the pandemic. An "infodemic," according to [12], is an excessive amount of information both online and offline that represents the spreading of wrong or unbased information. The WHO, alongside the pandemic, now had to release instructions about the growing problem of the unexpected "infodemic" of misinformation and disinformation. [13]

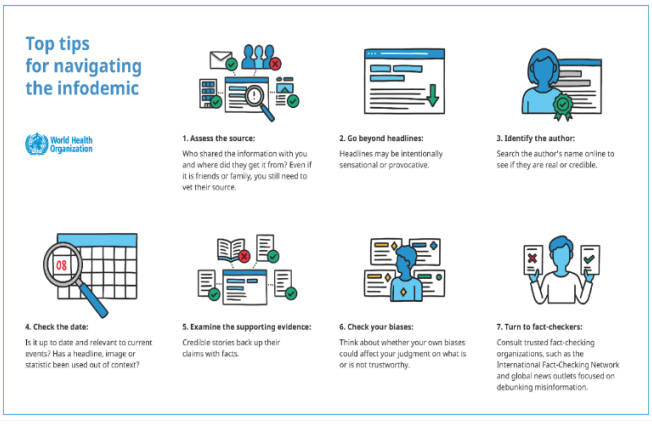

Figure 4. "Let's flatten the infodemic curve" issued by WHO about misinformation

Source: [13]

3.3. Hypotheses

For the purpose of this work, presented below are the hypotheses which throughout elaboration of this Thesis will be proven in their completion or non-completion. The Hypotheses for this Thesis are listed below:

H1. It is possible to build a tool that will be able to collect data from News Distribution sites.

H2. Unstructured data generated in H1 can be annotated using manual or automatic techniques.

3.4. Research Methodology

To research and study this problem comprehensively, we will use the following research methods:

- To prove the first hypothesis, we aim to build a tool for crawling data and periodically gather data from news outlets in North Macedonia as well as from various sources from the other Balkan countries.
- It is needed to identify important entities in the unstructured textual data.

  An automatic approach would be to develop an algorithm based on models such as tf-idf, that is expected to reflect how important a word is to a document in a corpus.

4. Data, Data Curation, and Algorithms

4.1. Big Data

With the evolution of new technologies and the advancement of the Web, the immense quantity of data in various forms is promptly produced, and the amount of data and information grows drastically. Internet users are flooded with Data and Information, especially now when a lot more devices are connected to the Internet, but the main issue remains that it is not simple to identify valuable information in all this overload of data, according to [14]

As data is being shared at high speeds continuously, the volume of information increases. This fast growth rate of Big Data has generated many challenges such as the rapid growth of data, speed of transfers, diverse data, and security concerns. Even So, Big Data is still in its formative years and will continue to develop. [15]

The definition of Big Data is based on the three letters V, which are: Volume, Velocity, and Variety.

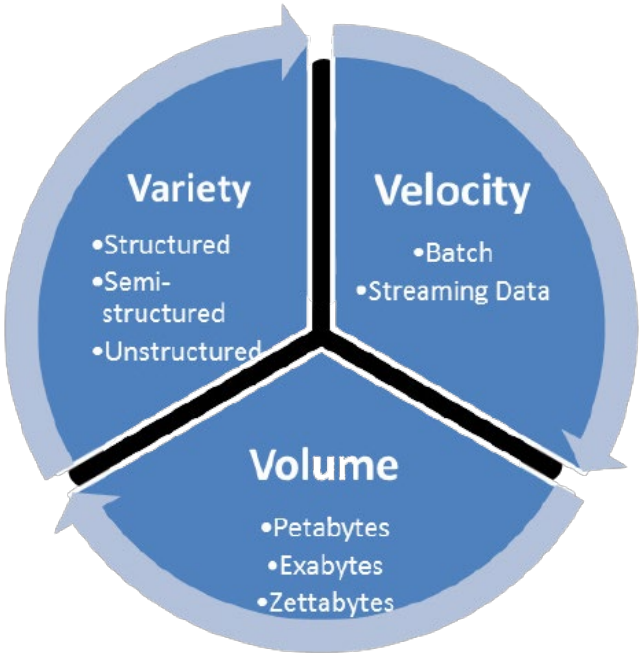

Figure 5. Big Data characteristics

Source: [16]

Volume describes the high volumes of unstructured data with uncertain value, such as Social Media posts and photos. Velocity stands for the rate at which the information is transferred or being shared, whereas Variety is referring to a variety of different data and information that is published. [17]

4.2. Data Curation

Data Curation is the process of selecting, collecting, maintaining, and archiving digital data information. This process maintains and gives worth, quality, and value to the data for it to become reusable for present and future use. [18]

Organizations and big Companies invest profoundly in Data analysis. However, most of them use only a minimal amount of the gathered data. Considering the big growth of data, together with the variety and its velocity, the process of getting the exact needed data for analysis is costly and time-consuming. Therefore, the process of Data Curation has a crucial role.

Data Curation has become important because of the enormous amount of data being shared on the Internet every moment. By Data Curation, it gets organized and managed and most importantly gets taken care of in quality hence making the Data provided trustful and easily retrievable for any kind of reuse. [19]

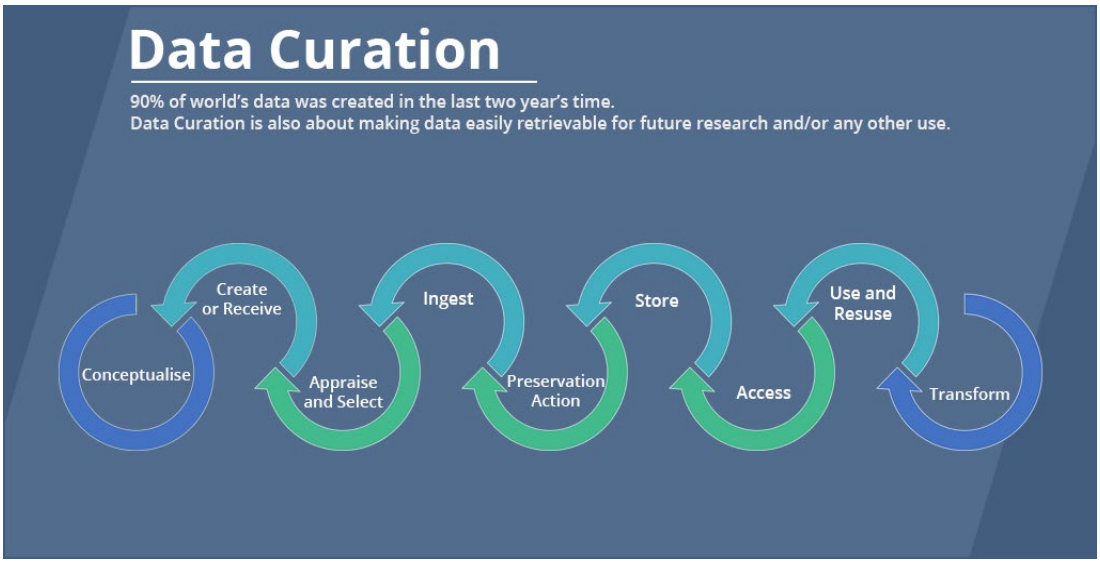

Figure 6. How Data Curation Works

Source: [20]

4.3. Data Clustering and Algorithms

4.3.1. What is Data Clustering?

Data Clustering is the process of grouping a set of data points into groups in which data points of a same group are more similar to one another than those on the other grouped data points, where the similarity measure is characterized in advance. [21]

Clustering is executed to obtain knowledge from data that is voluminous, which would be difficult to be analyzed from humans. Thus, Clustering Algorithms have developed as tools for meta knowledge gathering to operate exploratory data analysis. [22]

As explained in [23], although Clustering can be performed for different uses, below are listed some of the use cases in which Data Clustering can be applied for unsupervised learning in the Information Technologies subject:

- Fake News identification

Algorithms can work by taking over the content of the news article, analyzing the words and sentences, and clustering them. The clusters here are what helps the Algorithm decide which pieces are legitimate and which are fake news by looking into what words are more frequently used in eye-catching sensational articles. In this case, if the analysis shows a high percentage of specific words, it might be an indication of the news being fake. [23]

- Spam Filtering

The Spam or Junk Folder in our Email inboxes is another instance of Clustering and unsupervised learning.

Spam Filtering works by analyzing important sections of the email, like the sender, title, and content, and grouping them together. These groups are later classified to determine which emails are Spam. Clustering in the Classification process increases the accuracy to very high levels. [23]

- Website traffic Classification

In this case, a Webmaster wants to understand the traffic the website is getting to prevent spam traffic from bots. Algorithms are used here to cluster varieties from traffic sources.

After the clusters are created, the Webmaster is then able to classify traffic types according to his needs. By having precise information on the traffic, the Webmaster is able to grow the site and plan the work efficiently.

These were only some of the cases, more related to our subject, but Clustering and unsupervised learning can also be used in cases like Marketing and Sales, Document Analysis, Identifying illegal activities, Fantasy Football tactic games, and more, according to [23]

4.3.2. Algorithms and Techniques

According to [24], there are many Data Clustering Algorithms and methods that have become apparent as a powerful tool for meta-learning to precisely analyze the enormous amounts of data produced in the present times.

These Algorithms are mainly divided into five different types:

- Partitioning-Based
- Hierarchical-Based
- Density-Based
- Grid-Based
- Model-Based

4.3.2.1. Partitioning-Based

As explained in [24], in this type of Algorithm, the entire group of clusters is established promptly where Primary groups are specified and redistributed to a union.

Also explained as data objects being divided into some partitions by the partitioning algorithms, in which every partition characterizes a cluster.

These clusters must meet certain mandatory conditions, like Every group having at least one object, and each one of the objects must be affiliated to no more than only one group.

4.3.2.2. Hierarchical-Based

In this type of algorithms, data is organized in a hierarchical approach dependent on the standard of proximity, in which Intermediate nodes provide the Proximities.

Data sets are represented by a Dendrogram, and certain data gets presented by leaf nodes. The initial cluster steadily splits into a number of clusters as the hierarchy continues and progresses. [24]

4.3.2.3. Density-Based

As explained in [25], in this type of Algorithm, a cluster is a set of data objects that has distributed in the data space over a bordering area of high density of objects. Clusters in Density-based algorithms are split from one another by adjacent regions of low density of objects. Data objects located in such regions are usually deemed as noise or outliers.

4.3.2.4. Grid-Based

According to [24], Grid Based clustering algorithms are convenient in mining data sets that are sizeable and multidimensional.

The process here is applied by partitioning the data space into a finite number of cells that shape a grid structure. Then clusters are formed from the cells in the grid structure. Clusters relate to areas that are denser in data points than their surroundings. [24]

4.3.2.5. Model-Based

As explained in [26], Model-based Clustering is a common approach for high-dimensional data, which are more and more common, the model-based clustering approach has adapted to deal with it.

High-dimensional data have become in the present times more and more common but, unfortunately, classical model-based clustering algorithms have a disappointing performance in high-dimensional spaces.
Model-based clustering techniques are considerably over-parametrized.

All the above-mentioned techniques of Clustering Algorithm Methodologies under them include and cover many different, more intentional direct-approached Algorithms such as seen in Figure 7.

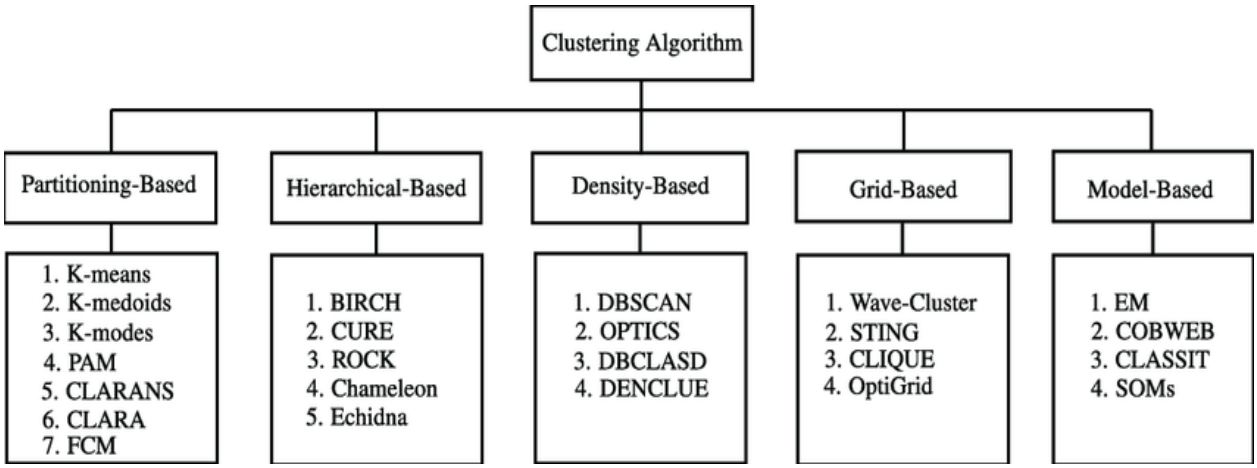

Figure 7. An overview of clustering taxonomy.

Source: [24]

Although all the above-mentioned Methods and Algorithms are very functional and available for Clustering in numerous use cases, the main ones covered in this Thesis will be Partitioning and Hierarchical-Based Algorithms.

4.4. Partitioning-Based Algorithms

4.4.1. K-means

According to [27], K-means is one of the most used clustering algorithms. K-means has been discovered, rediscovered, and studied by many researchers, scientists, and data engineers.

K-means is an iterative algorithm that works by having its data partitioned into prearranged groupings or clusters, where Euclidean or cosine could be used as distance measures. [27]

As explained in [28], the objective of K-means as an algorithm is straightforward: grouping of similar datapoints together to discover underlying patterns. For this objective, K-means searches for a fixed number (k) of clusters in a dataset.

A target number k is defined, which refers to the number of centroids needed in the dataset. A centroid is a location signifying the center of the cluster.

Each data point is distributed to each of the clusters via reduction of the in-cluster sum of squares. Namely, the K-means algorithm identifies k number of centroids, later allocates each data point to the nearest cluster. During this process, it tries to keep the centroids as small as possible. [28]

4.4.1.1. How K-means works

According to [29] [30], A well-structured workflow of the K-means algorithm would be as follows below:

1) Specifying the desired number of clusters K.

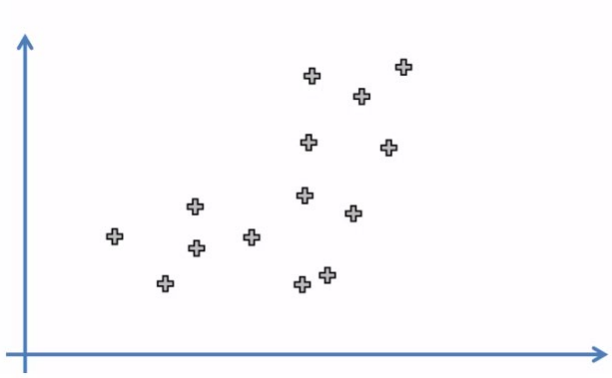

Figure 8. Step 1. Choosing the number of K Clusters (K = 2)

Source: [30]

2) Initializing centroids by initially shuffling the dataset, followed by randomly selecting K data points for the centroids without substitute.

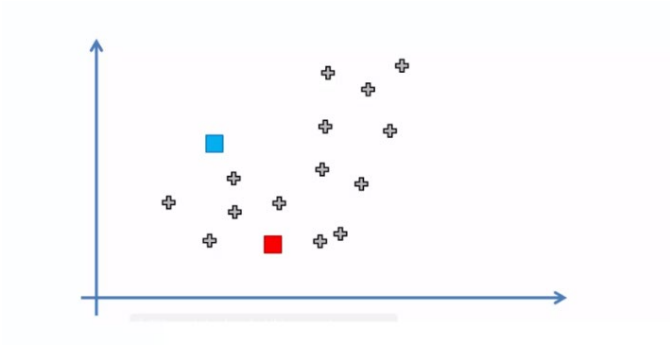

Figure 9. Step 2. Select at random K points, the centroids

Source: [30]

3) Continue iterating up until there is no change to the centroids.

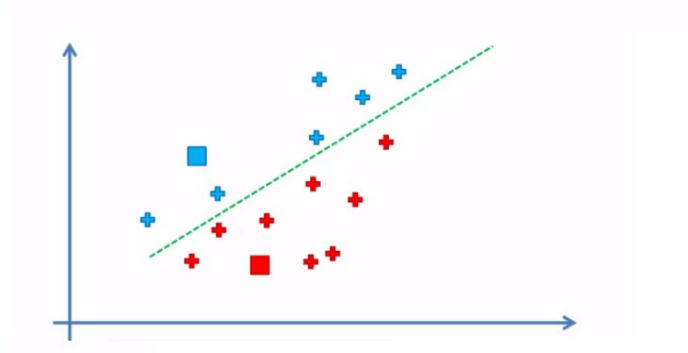

Figure 10. Step 3. Assigning each data-point to the closest centroid

Source: [30]

4) Computing the sum of the squared distance among the data points and all centroids.

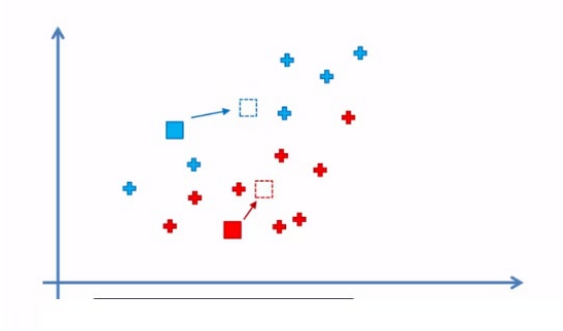

Figure 11. Step 4. Computing and placing the new centroid of each cluster

Source: [30]

5) Assigning of every data point to the nearest centroid.

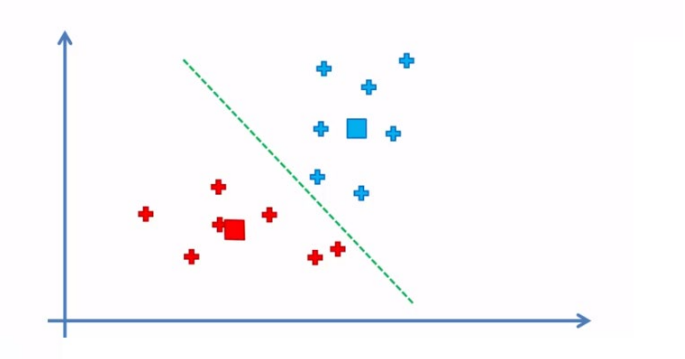

Figure 12. Step 5. Reassigning each data-point to the new closest centroid

Source: [30]

4.4.1.2. K-means Work Summary

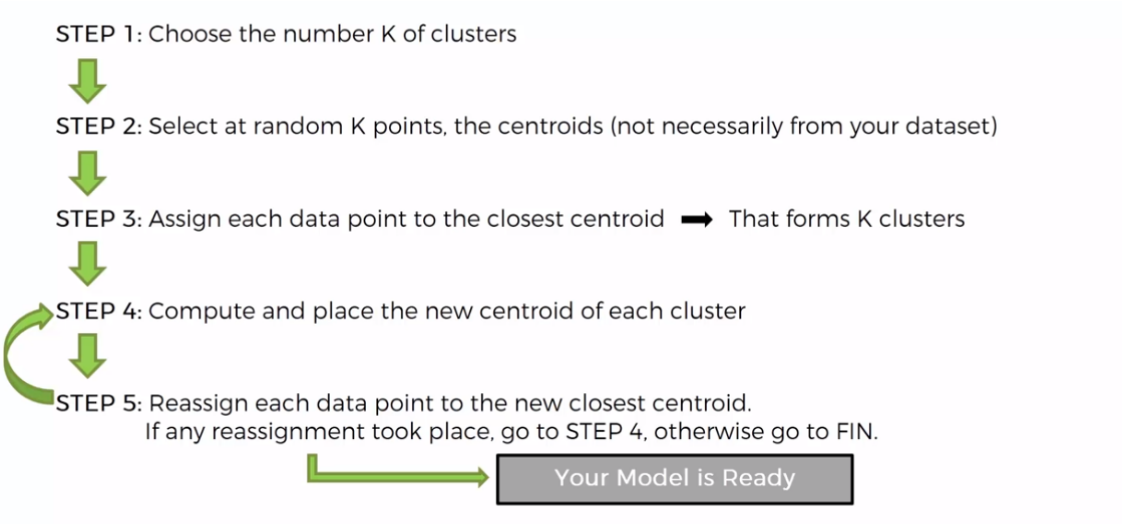

Figure 13. Step Summary of K-means work

Source: [30]

4.4.2. K-medoids

As explained in [27], the K-medoids algorithm applies partitioning around the medoids. Dissimilar to the k-means algorithm methodology, a medoid here represents any cluster. The distinctive object named "medoid" represents the most centrally positioned point inside the cluster.

Characteristic of the K-means algorithm is that it finds the mean to define the accurate center of the cluster, which may result in extreme values, whereas k-medoid calculates the cluster center using an actual point. The focus of this Algorithm is on its attempt to minimize the average divergence of objects against the object nearest to them. [27]

4.4.2.1. How K-medoids Algorithm works

According to [27] [28], the K-medoids clustering algorithm is named Partitioning Around Medoids (PAM), and these are the steps to follow for utilizing it:

1. Initialization

In the Initialization step, firstly, a random k is selected of the n data points as the medoid.

2. Assignment

Every data point must be linked to the nearest medoid.

3. Update centroids

In the case of having m-point in a cluster, swap the centroid prior with all additional (m-1) points from the cluster and complete the point as a new centroid that has a minimal loss.

4. Repeat

In the final step, repeat steps 2 and 3 up until convergence is reached.

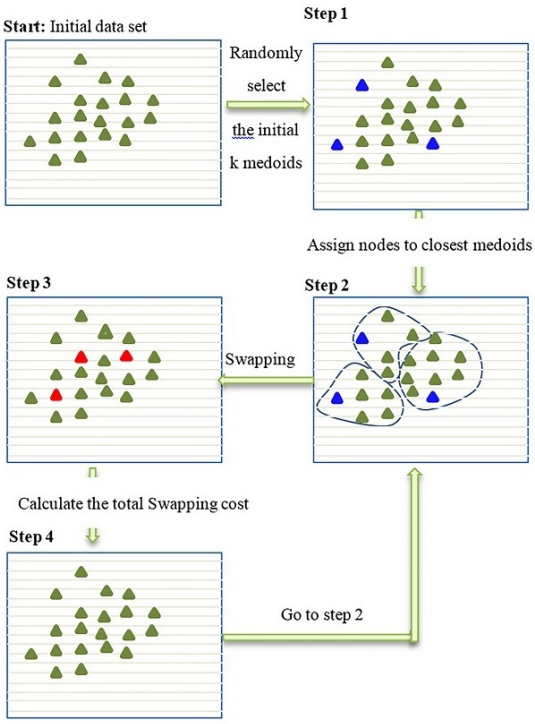

Figure 14. The k-medoids process

Source: [31]

4.5. Hierarchical-Based Algorithms

As analyzed in [32], Hierarchical Clustering can be described as a common technique used for unsupervised data analysis. It is a clustering tool used omnipresent in information retrieval, data mining, and machine learning.

Hierarchical-based techniques for Clustering correspond to a given dataset as a binary tree where every leaf represents a specific data point, but also each internal node represents a cluster on the leaves of its descendants.
HC as a clustering technique is most popular in fields such as analysis of Social Networks, bioinformatics, photo and text classification, and economic markets. [32]

According to [33], as an Agglomerative method, hierarchical Clustering is different from partition-based Clustering as it builds a binary merge tree beginning from leaves that hold data elements to the root that has the full dataset.

For implementing an HC algorithm, it is needed to choose a linkage function that specifies the distance among any two sub-sets and depends on the base space between the elements. An HC technique is monotonous only if the similarity drops along the way from any leaf to the root, or else there is at least one inversion. [33]

4.5.1. Types of Hierarchical Clustering

- Agglomerative Hierarchical Clustering
- Divisive Hierarchical Clustering

According to [34], Agglomerative Hierarchical Clustering is most commonly known as a bottom-up approach, in which each data or observation is considered as its cluster.

A duo of clusters gets combined until the situation where all clusters are merged in a large cluster that holds all of the data.

As explained in [34], Divisive Hierarchical Clustering, also called a top-down method. With this method, complete data or observation gets allocated to a specific cluster.

The cluster then gets split up until there is achieved one cluster for every single observation.

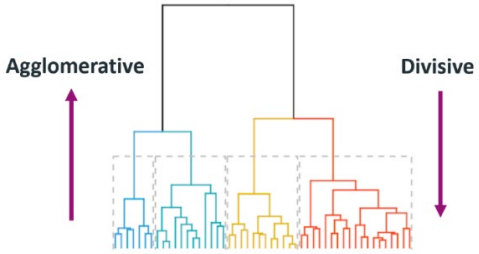

Figure 15. Agglomerative vs. Divisive approach

Source: [34]

Divisive Hierarchical Clustering is a rarely used approach, and we will not be covering it in this research further.

4.5.2. How Agglomerative hierarchical Clustering works [34]

1. Beginning with the assignment of every observation as a single cluster, in such a manner that if there are N observations, there are also N clusters, each one holding only one observation.

2. Locating the most similar and nearest couple of clusters and get them into one cluster, now there are N-1 clusters.

3. Locating the two closest clusters and making them into one cluster. There are now N-2 clusters. This is accomplished using the method of agglomerative clustering linkage.

4. Repetition again of the above, steps 2 and 3 up until achieving all observations being clustered into one distinct cluster of size N.

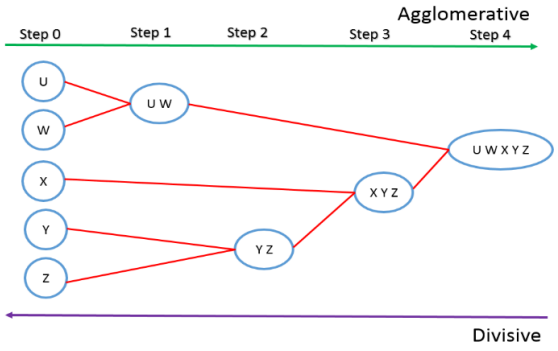

Figure 16. How Agglomerative Hierarchical Clustering Algorithm Works

Source: [34]

4.5.3. Distance or Dissimilarity methods

Clustering algorithms utilize several distance or dissimilarity methods to create different clusters. Lower distance suggests data or observations are similar and could be put together in a single cluster.

To identify similar and dissimilar measures, Step 2 of Agglomerative hierarchical Clustering may be performed in several approaches: [34]

1) Euclidean Distance

   Euclidean Distance is a constant metric that, in geometric terms, is defined as the straight-line distance in the middle of two points.

   The Euclidean distance among two points determines the length of a segment connecting the two points. Euclidean Distance is the most obvious method of demonstrating the distance between two points.

   For calculation of the distance between two points, the Pythagorean Theorem is used, as shown in the figure below. In case the points (x1, y1) and (x2, y2) are in a two-dimensional surface, the Euclidean distance in between is as seen in Figure 17.

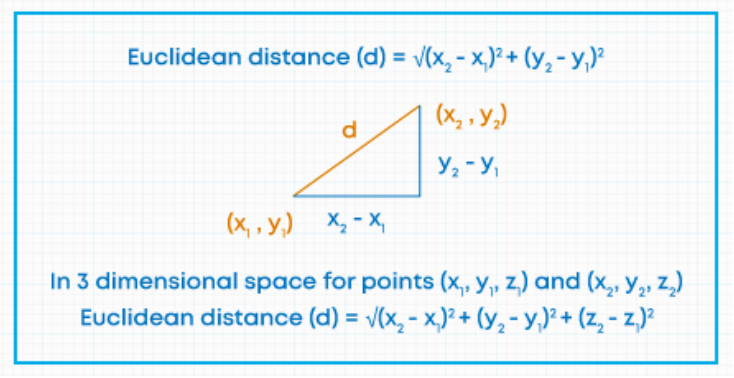

Figure 17. Euclidean Distance illustration

Source: [34]

2) Manhattan Distance

   The distance measured among two spots appearing in a grid-based, on a rigorously horizontal and vertical way. Manhattan distance is the plain sum of the vertical and horizontal pieces.

Manhattan distance can also be described as the distance if travel only along the coordinates had to be made.

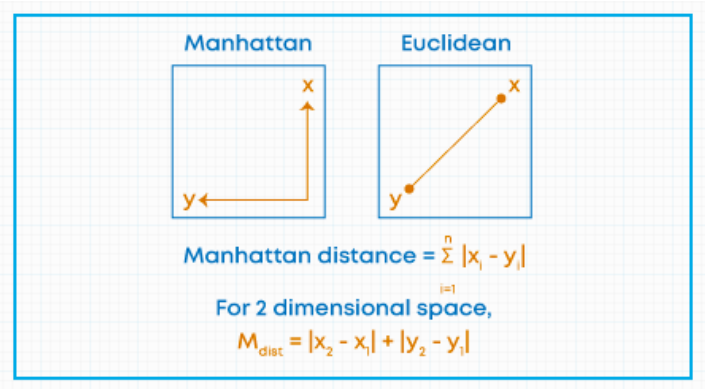

Figure 18. Manhattan Distance illustration

Source: [34]

3) Minkowski Distance

Minkowski distance represents a measurement of distance or similarity for two points in a normed vector space. It is a generalization of the above-mentioned Euclidean and Manhattan distance approaches.

The distance between the variables X and Y is described as shown below.

$$\left(\sum_{i=1}^{n} |X_i - Y_i|^p\right)^{1/p}$$

If p = 1, it is corresponding to the Manhattan distance, but if p = 2, it is corresponding to the Euclidean distance.

4) Jaccard Similarity Coefficient

As stated in [35], Jaccard Similarity is a statistic used for measuring the similarity and dissimilarity of sample sets.

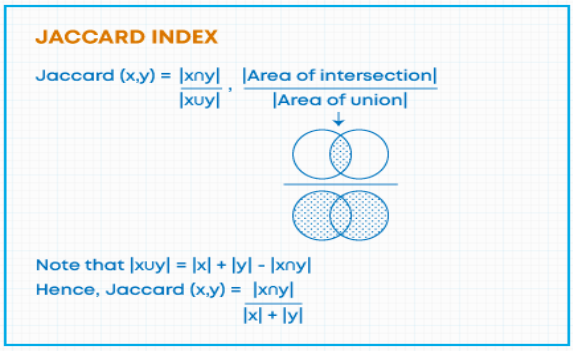

Figure 19. Jaccard Index

Source: [34]

Jaccard Similarity is a very powerful formula used in cases such as: “object detection in image recognition, classification, and image segmentation tasks."

This approach is utilized mostly when the data or variables are of a qualitative nature. Used when the variables are signified in binary form.

5) Cosine Similarity

As described in [36], Cosine similarity represents a metric applied to define the similarity of documents, irrespective of their size.

This formula measures the cosine of the angle between two vectors estimated in a space of multi-dimensions. Within this frame of reference, the two vectors stand as arrays holding the word counts of two documents.

As shown below, A and B are two vectors for comparison. Utilizing the cosine measurement as a similarity function

$$\text{similarity (A, B)} = \frac{A.B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

6) Gower’s Similarity Coefficient

Gower’s Similarity Coefficient as an approach is used for varied data types, namely databases, mostly in cases when data covers both qualitative and quantitative variables.

4.5.4. Finding the Optimal number of Clustering

A challenging task in agglomerative Clustering is the process of finding the optimal number of clusters. In such tasks, Silhouette Score is a popular approach to measure how close every point inside a cluster stands to the points in its nearby clusters. [34]

$$S_i = \frac{b_i - a_i}{Max(a_i, b_i)}$$

4.6. TF-IDF

TF-IDF, in short, means "Term Frequency — Inverse Document Frequency." It is a statistical technique that is utilized in information retrieval, text mining, and for purposes of identifying how relevant a term or a keyword is to a given document. [37]

With the quick expansion and advancement of the Internet, data and information grew rapidly as well. Enormous amounts of text data need to be filtered out in order to provide users the information they need. In such a condition, automatic text classification is a very helpful assistance. [38]

4.6.1. Term Frequency (TF)

TF is used for measuring how frequently a word is present in a document. This is also highly impacted by the length of the document and the generality of words.
Also defined as the ratio of a word's presence inside a document to the overall amount of words. [39]

$$\text{TF (w, d)} = \frac{occurrences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

A simple example here would if we have 2 Documents, namely document A and document B, with the following present words, as shown in Figure 20.

Documents	Text	Total number of words in a document
A	**Usa is the best country**	5
B	**Russia is the biggest country in the world**	8

Figure 20. Example of Term Frequency

As explained in [39], The primary action step is to make a vocabulary of unique phrases and evaluate Term Frequency for every document. Term Frequency is relevant more for words that commonly are found in a document and less for uncommon terms.

Words	TF for document A	TF for document B
USA	1/5	0
Is	1/5	1/8
The	1/5	2/8
Best	1/5	0
Country	1/5	1/8
Russia	0	1/8
Biggest	0	1/8
In	0	1/8
World	0	1/8

Figure 21. Term Frequency Result for Document A and B

4.6.2. Inverse Document Frequency (IDF)

According to [39], Inverse Document Frequency represents the measure for the importance of a term. Term Frequency (TF) does not contemplate the significance of keywords.

In many cases, most present are commonly used terms like "of," "and" etc., but these terms do not really represent value.

Inverse Document Frequency gives weightage for every word based on the word's frequency in the corpus D. IDF formula is described as shown below.

$$\text{IDF (w, D)} = \ln\left(\frac{Total\ number\ of\ documents\ (N) in\ corpus\ D}{number\ of\ documents\ containing\ w}\right)$$

An example here would be as shown in Figure 22, where present are documents A and B in the corpus, N=2.

Words	TF for document A	TF for ducment B	IDF
USA	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
Best	1/5	0	ln(2/1) = 0.69
Country	1/5	1/8	ln(2/2) = 0
Russia	0	1/8	ln(2/1) = 0.69
Biggest	0	1/8	ln(2/1) = 0.69
In	0	1/8	ln(2/1) = 0.69
World	0	1/8	ln(2/1) = 0.69

Figure 22. IDF example

4.6.3. Term Frequency + Inverse Document Frequency (TFIDF)

TFIDF, the result of TF and IDF.

TFIDF applies further weightage for the term that is not very common in the corpus (all the documents). TFIDF gives more significance to the term that is more common in the document. [39]

$$TFIDF\ (w, d, D) = TF\ (w, d) * IDF\ (w, D)$$

Following the utilization of TFIDF, text in documents A and B can be exemplified as a "TFIDF vector of dimension" equivalent to the vocabulary phrases. The related importance for every word characterizes the significance of that word in a certain document. [39]

Words	TF for document A	TF for document B	IDF	TFIDF for document A	TFIDF for document B
USA	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
Best	1/5	0	ln(2/1) = 0.69	0.138	0
Country	1/5	1/8	ln(2/2) = 0	0.138	0
Russia	0	1/8	ln(2/1) = 0.69	0	0.086
Biggest	0	1/8	ln(2/1) = 0.69	0	0.086
In	0	1/8	ln(2/1) = 0.69	0	0.086
World	0	1/8	ln(2/1) = 0.69	0	0.086

Figure 23. TFIDF for Documents A and B

Term Frequency — Inverse Document Frequency (TFIDF) overall represents a good method for text vectorization centered on the BoW (Bag of Words) standard. The major disadvantage here is that it does not recognize the semantic meaning of phrases, but this drawback is easily overcome by further sophisticated methods such as word2Vec. [39]

4.7. Topic Modelling

Topic Modeling is a technique used for data clustering that represents statistical modeling for discovering and identifying topics that occur in a set of documents and obtaining hidden patterns exposed by a text amount, therefore, helping in improved decision making. [40]

According to [40], Topic Modelling is distinct compared to other rule-based text clustering approaches that utilize dictionary word searching or use regular expressions. Topic Modelling is an unsupervised method applied for obtaining and observing the collection of terms in significant clusters of text data that are signified as topics.

According to [41], At the document level, one of the most effective approaches for text understanding is by exploring its topics. The main idea for all topic models is represented as follows:

- Documents contain a mixture of topics
- Topics comprise of a collection of words

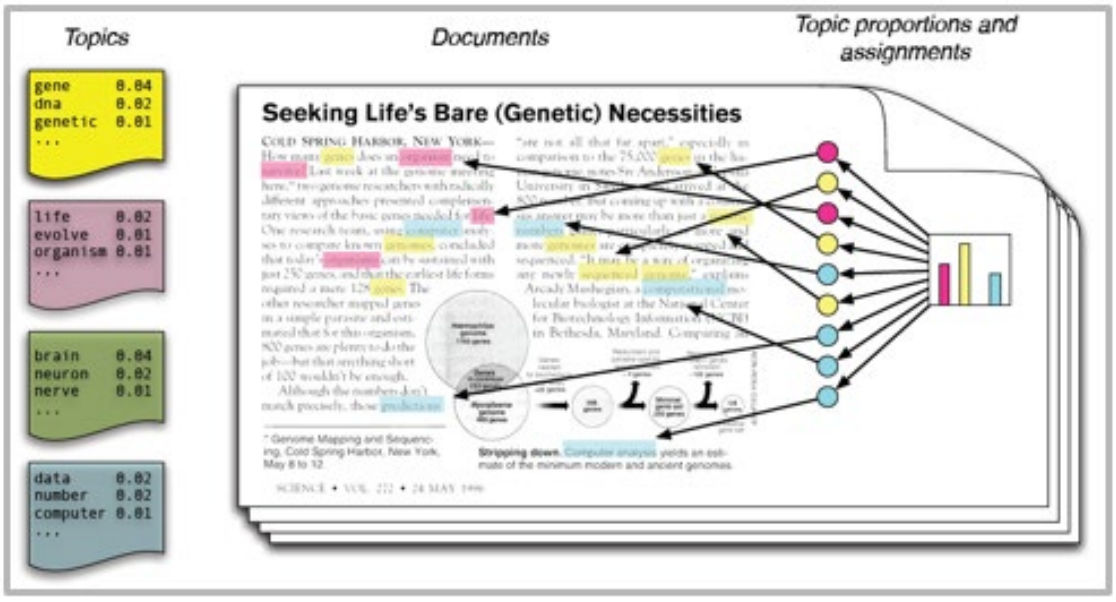

Figure 24. Topic Modelling of a Document

Source: [40]

Certain well-known methods for topic modeling are listed below:

- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)
- Non-Negative Matrix Factorization (NMF)

4.7.1. LSA - Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique that is represented as one of the foundational techniques in topic modeling. The main theory is to take a matrix from the targeted document and decompose it into a distinct document-topic matrix and a topic-term matrix applying Singular value decomposition. [41]

According to [41], in Latent Semantic Analysis, Rows signify words whereas columns mean documents. It is usually applied as a noise-reducing technique, otherwise called dimension reduction.

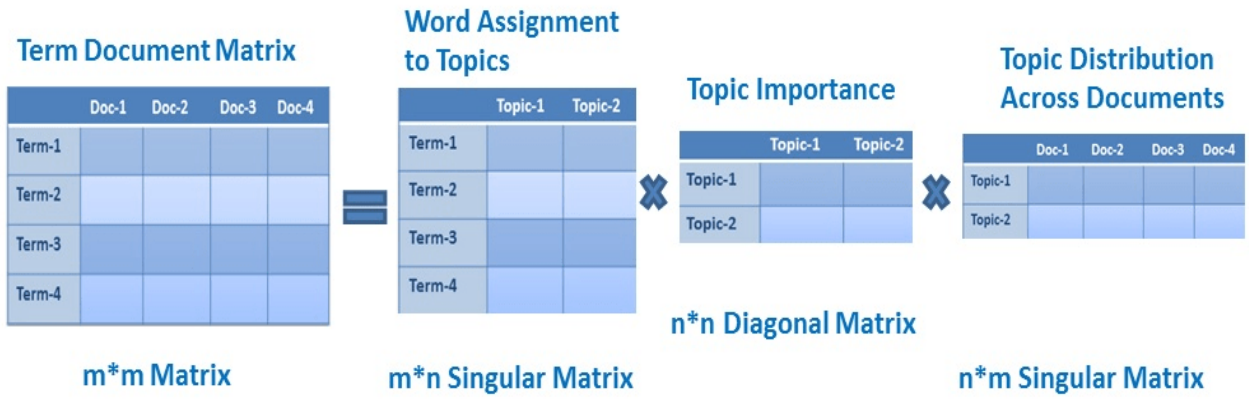

Figure 25. LSA - Latent Semantic Analysis

Source: [41]

4.7.1.1. Singular Value Decomposition

Singular Value Decomposition is a technique for matrix factorization which exemplifies a matrix in the product of two matrices. [41]

The formula for SVD is shown below:

$$M=U\sum V^*$$

This formula is further clarified as:

- Letter “M” represents an m×m matrix
- Letter “U” represents a m×n left singular matrix
- Sigma “Σ” represents a n×n diagonal matrix by non-negative real numbers.
- Letter “V” represents a m×n right singular matrix
- V* represents n×m matrix, which signifies the transpose of V.

4.7.2. LDA - Latent Dirichlet Allocation (LDA)

According to [41], Latent Dirichlet allocation (LDA) represents a generative paradigm that enables sets of observations to be described by unobserved groups that explain why certain parts of the data are similar and alike. LDA has had a significant influence in spheres like natural language processing and statistical machine learning; thus, it has become a very popular probabilistic modeling technique in machine learning.

As explained in [42], “Latent Dirichlet Allocation (LDA) model is innovatively developed to understand latent driving states and quantified structure of the driving behavior patterns (topics) from individualization driving (documents) using driving behaviors (words).”

In LDA, documents display numerous topics. In text pre-processing, punctuations and stop phrases (such as "if," "the," or "on," which contain little topical content) are excluded. Thus, every document is considered as a combination of corpus-wide topics. A topic is a distribution over a static vocabulary, and topics get produced from collections of documents. [43]

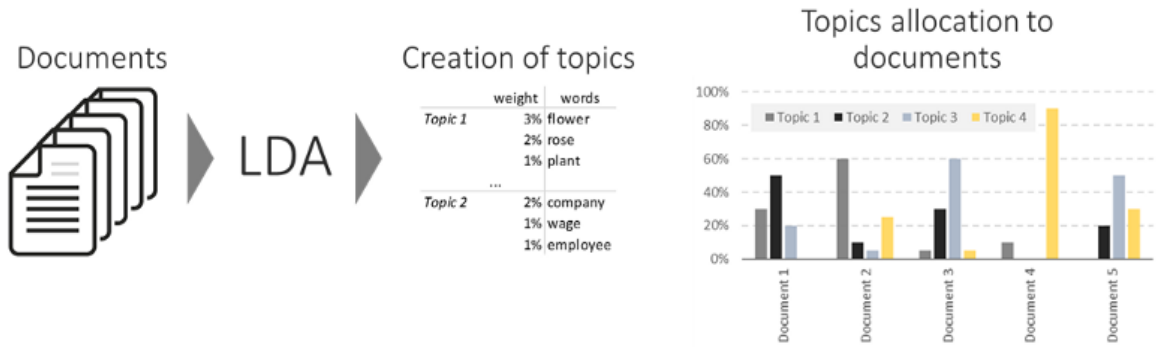

Figure 26. Illustration of LDA input/output workflow

Source: [44]

As explained further in [44], there are examples where we could have words like “flower" and "rose" with high probability. Documents have a probability distribution over topics, in which every phrase is considered as derived from one of those topics. After such distribution of document probability over topics, the analysis concludes with results of how much a topic is implicated in a document.

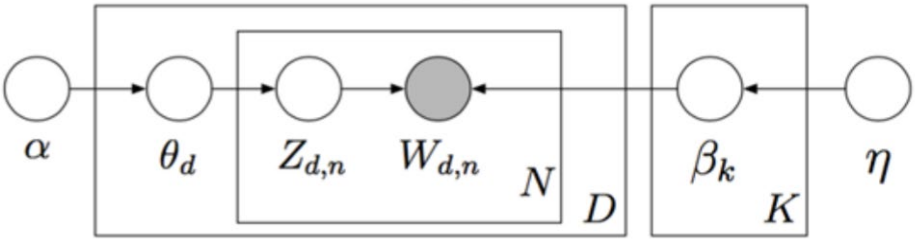

Figure 27. Graphic model for Latent Dirichlet allocation

Source: [43]

Figure 27 explanation, from source [43]:

Initially, α is a proportion parameter and η a topic parameter. Topics are represented as $\beta_{1:K}$, where every β_k is a dispersal over the vocabulary.

The topic proportion for the d document is θ_d, in which $\theta_{d,k}$ represents the topic proportion for topic k in document d. The topic assignments for document d, are Z_d, in which $Z_{d,n}$ is the topic assignment for word n in document d.

Lastly, the remarked words for document d are w_d, in which $w_{d,n}$ is the word n in document d, that represents an element taken the fixed vocabulary.

4.7.3. NMF – Non-Negative Matrix Factorization

Non-Negative Matrix Factorization can be described as a statistical technique used to minimize the dimension of the input corpora. NMF applies the factor analysis method to deliver comparatively less weightage to phrases that have less coherence. [45]

According to [46], "NMF takes advantage of the fact that the vectors are non-negative". NMF uses factorization for vectors in the lower-dimensional manner and forces coefficients to be non-negative to themselves.

$$\frac{1}{2} || \mathbf{A} - \mathbf{WH} ||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (A_{ij} - (WH)_{ij})^2$$

A further detailed explanation for this equation could be as follows:

Having the initial matrix, A, obtainable are two matrices, namely W and H, thus A= WH. As explained in [46], NMF has a property of inherent Clustering, meaning that W and H signify the subsequent information below about A:

- A (Document-word matrix) — input that holds the terms that occur in certain documents.
- W (Basis vectors) — the topics obtained from the documents.
- H (Coefficient matrix) —affiliation weights about the topics in the documents.

W and H are calculated by enhancing an objective function (such as the EM algorithm), getting both W and H updated iteratively until convergence is reached.

As shown in the equation above, measured is the error of reconstruction among A and the product of A's factors, namely W and H. The process is established on Euclidean distance. [46]

By utilizing the objective function, updating rules for W and H are obtained, and the results are as shown in the equations below.

Updated values for W

$$W_{ic} \leftarrow W_{ic} \frac{(\mathbf{A}H)_{ic}}{(\mathbf{WH}H)_{ic}}$$

Updated values for H

$$H_{cj} \leftarrow H_{cj} \frac{(W\mathbf{A})_{cj}}{(W\mathbf{WH})_{cj}}$$

The values that got updated get calculated in parallel operations. By using the newly updated W and H, the reconstruction error is re-calculated repeatedly until convergence is achieved. [46]

5. Crawling Data, Designing and populating the dataset

For the purpose of this work, we aim to collect our data from news sites in Albanian and organize them into Positive data and Negative data.

Positive data represent fake news, while negative data are news articles that come from valid sources. All articles are in the Albanian language.

We have created a list of media from which we collect data. Since most fake news appears in politics, we have focused on that topic. The media that we collect data are listed in the following table:

Media	URL	Publishes Fake News
Tetova1	https://tetova1.com/feed/	No
Lapsi.al	https://lapsi.al/feed/	No
Fakti Ditor	http://www.faktiditor.ch/feed/	No
Illyrian News Agency - INA	https://ina-online.net/feed/	No
Medial.mk	https://medial.mk/feed/	No
ABNTimes24.com	https://abntimes24.com/feed/	Yes
NewsB99.com	https://newsb99.com/feed/	Yes
RTN24News.com	http://rtn24news.com/feed/	Yes
WorldNetPress.com	https://worldnetpress.com/feed/	Yes
KosAlb.net	https://kosalb.net/feed/	Yes

We notice that most of the Fake News sites use unusual and meaningless domain names for them to be in the Albanian language. According to our findings, this is because these domains are sold to Albanians, from other first-time creators, mainly from countries such as India, Russia, and other Asian countries.

The domain creators create them in mass with the purpose of selling them. The domains are firstly approved for Monetization services such as Facebook Audience Network, Google AdSense, or MGID, and only then are sold to the buyers.

The Albanian publishers of Fake News do not really consider the importance of a serious domain as their main audience does not come from direct domain visitors but rather from social media, mostly sharing in big Facebook Pages and Groups, where the domain name is overshadowed by the big clickbait Titles and eye-catching featured images.

For the sites that are Monetized with Facebook Audience Network, when a user clicks a News Article, it does not even open in a Browser because Facebook has its own In-app viewing in which it can display its own Ads; thus, the domain relevance is overlooked.

5.1. Crawling the News Articles

Almost all News sites in the Albanian language use the WordPress platform for posting news articles. Being a free, highly customizable, and open-source content management platform, it is very convenient in this field of News Distribution.

WordPress as an open-source offers many tools and options for Content management and also support by a very large community. [47]

5.1.1. RSS Feed

WordPress automatically includes a very useful utility for the clients called the RSS Feed. It is an easy way to access all the news articles posted by the News source.

RSS stands for “Really Simple Syndication”. It represents the information in XML as an easy-to-read format for computers and a very convenient way for news crawling in our use case. [48]

RSS Feeds are updated in real-time. The Feed includes each article as an <item> tag, and also the content in a formatted form such as <title>, <link>, <pubDate>, and all content paragraphs as <p>. These tags are what we will be looking for with our Crawling application.

The default URL for the RSS feed is “domain.com/feed”. RSS Feed uses Pagination, with one Page usually displaying 10 news articles. To continue to the next page, we need to add to the URL: “/feed/?paged=2” and so on. This way, we will crawl about 100 RSS feed pages.

5.2. RSS Feed Crawling – Windows application

We have developed a Windows program with the Python programming language that utilizes the RSS Feed to crawl the data for our chosen websites.

The application User Interface can be seen in Figure 28.

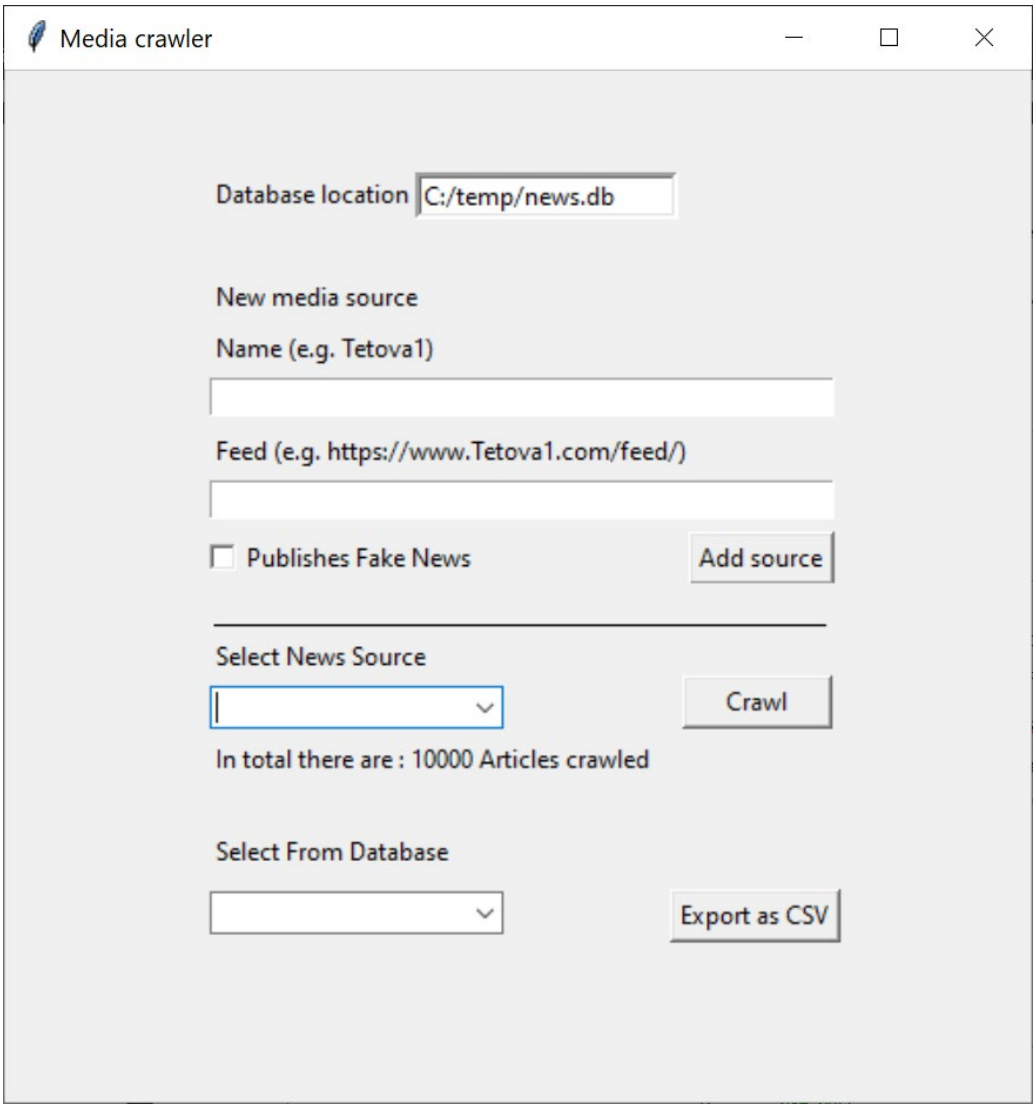

Figure 28. Media Crawler Windows application

We use “tkinter” package for our GUI, “requests” package for HTTP requests, “bs4” for getting data from XML files, and “pandas” package for Data Manipulation.

```
from tkinter import *
from tkinter.ttk import Combobox
from bs4 import BeautifulSoup
import sqlite3, time, csv
import pandas as pd
from sqlite3 import Error
import requests
import tkinter as tk
from tkinter import filedialog
from tkinter.filedialog import asksaveasfile
```

For storing the data, we have used SQLite as a simple database solution. The default database location is set to a safe location such as “C:/temp/news.db”.

For saving the News sources we have created a table called “domain” with the following structure: ***id:int, media:text, media_feed:text, isFake:int***

```
    sql_create_domain_table = """ CREATE TABLE IF NOT EXISTS domain (
                                        id integer PRIMARY KEY,
                                        media text NOT NULL,
                                        media_feed text,
                                        isFake integer
                                    ); """
```

For saving the News content, we have created two tables, one for Negative and one for Positive data. The data is saved in or the other database Table, depending on if the user checks the "Publishes Fake News“ option in the application, which invokes "isFake" in the Database. In case the user checks that option, the data is saved in Positive data. Data is saved in the Negative table if the user does not check the "Publishes Fake News" option.

The structure for Negative and Positive data is:

Id:int, media:text, media_feed:text, text_title:text, text_content:text, url:text, date_published:text

```
sql_create_positive_table = """ CREATE TABLE IF NOT EXISTS positive (
                                        id integer PRIMARY KEY,
                                        media text NOT NULL,
                                        media_feed text,
                                        text_title text,
                                        text_content text,
                                        `url` text,
                                        date_published datetime
                                    ); """

    sql_create_negative_table = """ CREATE TABLE IF NOT EXISTS negative (
                                        id integer PRIMARY KEY,
                                        media text NOT NULL,
                                        media_feed text,
                                        text_title text,
                                        text_content text,
                                        `url` text,
                                        date_published datetime
                                    ); """
```

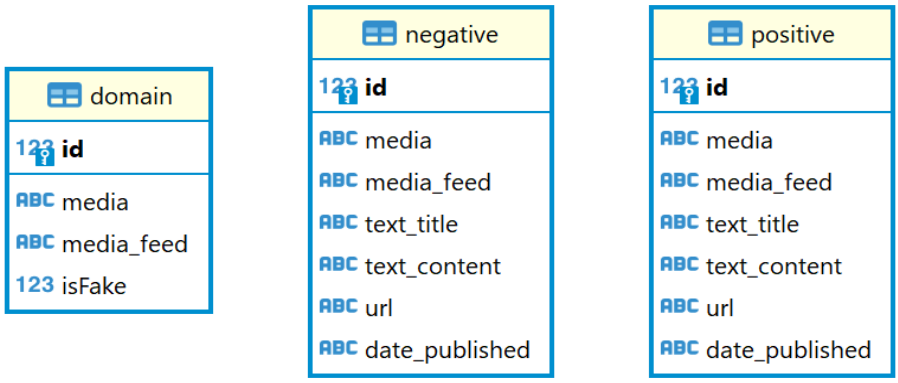

Figure 29. Database Diagram

5.2.1. Media Crawler – Functionality

The Crawling application that we have developed offers the user the option of Naming the News Media Source with a custom Name, followed by adding the RSS Feed URL of that News media source. In case the Media publishes Fake News, the user checks the checkbox "Publishes Fake News" and adds the Source to the database.

Figure 30. Adding new Media Source

```
def add(self):
        print("add btn")
        media = self.t22.get()
        media_feed = self.t33.get()
        isFake = self.chk_state.get()
        self.insert_data(media, media_feed, isFake)
        print(media, media_feed)
        self.t22.delete(0, 'end')
        self.t33.delete(0, 'end')
        media_tuple = self.select_all_feed()
        self.data = media_tuple
        # self.cb = Combobox(win, values=self.data)
        self.cb.config(values=self.data)
        self.cb.update_idletasks()
```

The Media sources that we have added to the database can be accessed in the Dropdown list offered under the “Select News Source” section of the application, as can be seen in Figure 31.

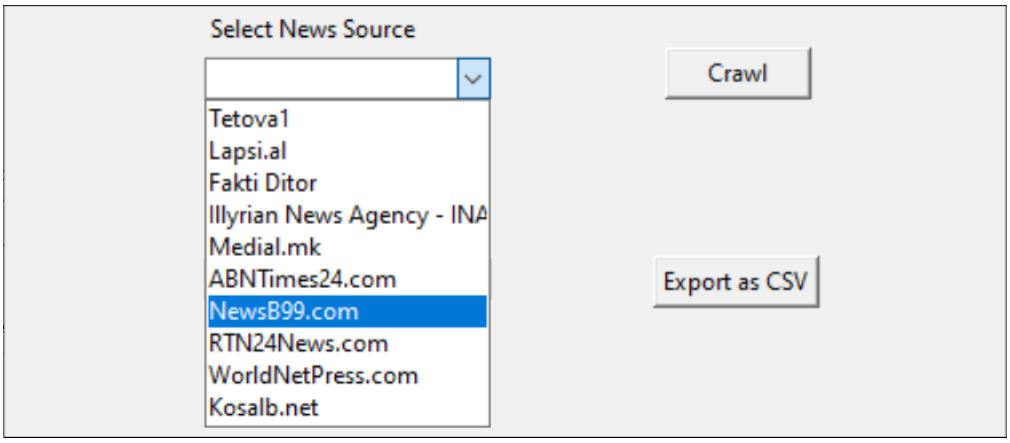

Figure 31. Added Media Sources Dropdown list

```
def select_all_feed(self):
        try: c = self.conn.cursor() # c.execute(create_table_sql)
            sql = "SELECT media FROM domain"
            c.execute(sql)
            rows = c.fetchall()
            media_tuple = ()
            y = list(media_tuple)
            for row in rows:
                y.append(row[0])
            media_tuple = tuple(y)
            return(media_tuple)
        except Error as e:
            print(e)

    def get_details(self, media):
        try: c = self.conn.cursor()
            sql = f"SELECT * FROM domain where media = '{media}'"
            c.execute(sql)
            rows = c.fetchall()
            return(rows)
        except Error as e:
            print(e)
```

After selecting a News Source from the Dropdown list, the user is ready to click the "Crawl" button, and the application starts crawling News Articles from the RSS Feed URL that has been inputted by the user for that exact News outlet.

The crawling process is executed as shown in Figure 32.

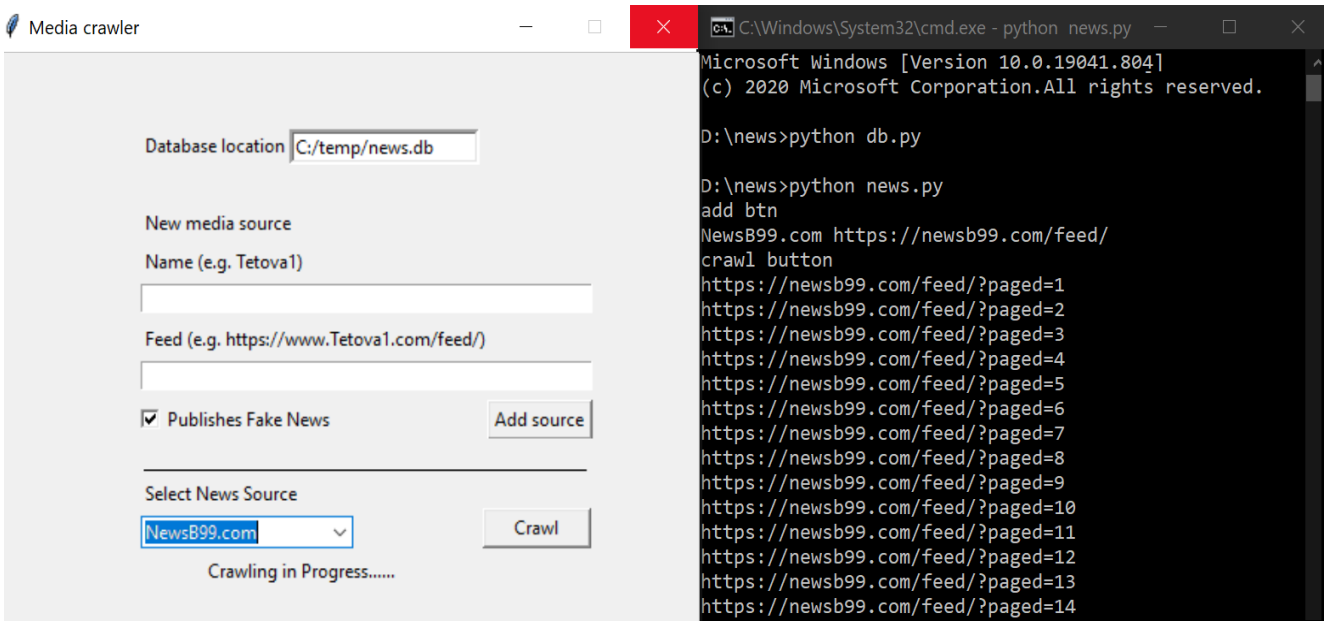

Figure 32. Crawling Process

```
    def scraping(self, media, isFake):
        url = self.main_url + "?paged="
        site_url = url

        for i in range(100):
            url = site_url + str(i + 1)
            print(url)
            r = requests.get(url)
            soup = BeautifulSoup(r.text, features='lxml')
            # print(soup.title)

            for a, item in enumerate(soup.select('item')):
                p_text = ""

                for p_tag in item.findAll("p"):
                    p_text += p_tag.text + "\n"
                # print (p_text)


                title = item.select_one("title").text
                link = item.link.next_sibling
                pubdata = item.select_one("pubDate").text
                p_text = p_text

                self.insert_crawled_data(media, title, link, pubdata, p_text, isFake)
```

All the crawled data by the application is then stored into the database either in Positive or Negative data Tables, by the following code

```
self.insert_crawled_data(media, title, link, pubdata, p_text, isFake)

    def insert_crawled_data(self, media, title, link, pubdata, p_text, isFake):
        try:
            c = self.conn.cursor()
            # c.execute(create_table_sql)
            if(isFake == 0):
                sql = "INSERT INTO positive (media, media_feed, text_title, text_content,
'url', date_published) VALUES ( ?, ?, ?, ?, ?, ?)"
            else:
                sql = "INSERT INTO negative (media, media_feed, text_title, text_content,
'url', date_published) VALUES ( ?, ?, ?, ?, ?, ?)"
            c.execute(sql, (media, self.main_url, title,  p_text, link, pubdata))
            self.conn.commit()
```

5.3. Exporting the datasets

The application we have developed offers the option to Export the data in a dataset format .CSV. The User firstly selects one of the two Database Tables that contain the News articles and can export a Dataset with news articles.

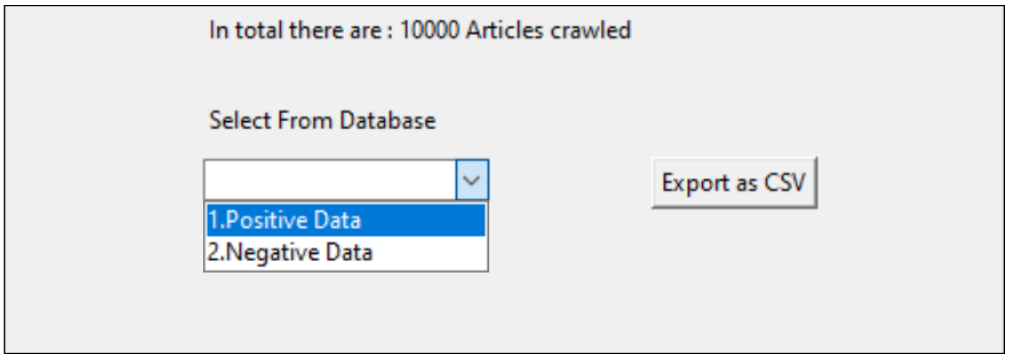

Figure 33. Exporting the Datasets

```
def download(self):

        files = [('CSV Files', '*.csv'),
            ('Text Document', '*.txt')]
     directory = filedialog.asksaveasfilename(filetypes = files, defaultextension = files)
        selected_txt = self.cb2.get() print(selected_txt)
        if (selected_txt == "1.Positive Data"):
            selected_db = "positive"
        elif(selected_txt == "2.Negative Data"):
            selected_db = "negative"
        new_conn = sqlite3.connect("C:/temp/news.db", isolation_level=None,
                                   detect_types=sqlite3.PARSE_COLNAMES)
        db_df = pd.read_sql_query(f"SELECT * FROM '{selected_db}'", new_conn)
        db_df.to_csv(directory, index=False, encoding='utf-8-sig')
```

After exporting the datasets as a .CSV file, an example of the data inside the dataset looks as seen in Figure 34.

	A	B	C	D	E	F	G
652	651	ABNTimes24.co	https://abntimes24.com/feed/	Qytetarët kanë mbetur të befasuar nga këto pamje, ja çfarë ka bërë Edi Rama që në mëngjes	Kryeministri i vendit Edi Rama e ka nisur ditën me shumë pozitivitet. Ai ka ndarë në rrjetet sociale një video me pamje nga Bulevardi i Ri të cilat kanë lënë gojëhapur qytetarët. Ai i ka uruar të gjithëve një javë të mbarë dhe nuk kanë qënë të paktë ata qytetarëve që kanë hamendësuar se këto pamje janë 3D dhe jo realitet. Po ju si mendoni?! Shtyp/Hap hapesiren e meposhtme qe te shikoni videon	https://abntimes24.com/qytetaret-kane-mbetur-te-befasuar-nga-keto-pamje-ja-cfare-ka-bere-edi-rama-qe-ne-mengjes/?utm_source=rss&utm_medium=rss&utm_campaign=qytetaret-kane-mbetur-te-befasuar-nga-keto-pamje-ja-cfare-ka-bere-edi-rama-qe-ne-mengjes	Mon, 17 May 2021 09:07:07
653	652	ABNTimes24.co	https://abntimes24.com/feed/	Alfred Cako zbulon përse Pandeli Majko mbeti pa mandat në zgjedhjet e 25 prillit: "Ka gisht Rama dhe..."	I ftuar ditën e djeshme në Top Channel, konspiracionisti Alfred Cako zbuloi për herë të parë përse mendon se Pandeli Majko mbeti pa mandatin e deputetit në këto zgjedhje. Ai tha se këtu kishte dorë Edi Rama, ishte diçka e qëllimshme pasi Majko ishte një ndër të preferuarit e amerikanëve. "Pandeli Majkoja ka qenë i preferuari i amerikanëve, është një njeri që hiqet sikur është parimor, por në fakt nuk është parimor e as naiv. Është shumë i sofistikuar. Ai është përpjekur si zvarranik duke treguar ndonjëherë dhëmbët po duke menduar se do t'i vinte ndonjëherë radha pas dështimit të Ramës. Por Edi Rama duke e ditur këtë avash avash e largoi si figurë nga PD-ja", tha Cako. Shtyp/Hap hapesiren e meposhtme qe te shikoni videon	https://abntimes24.com/alfred-cako-zbulon-perse-pandeli-majko-mbeti-pa-mandat-ne-zgjedhjet-e-25-prillit-ka-gisht-rama-dhe/?utm_source=rss&utm_medium=rss&utm_campaign=alfred-cako-zbulon-perse-pandeli-majko-mbeti-pa-mandat-ne-zgjedhjet-e-25-prillit-ka-gisht-rama-dhe	Mon, 17 May 2021 09:02:26

Figure 34. Example of how crawled news articles look inside a dataset

We have exported two datasets, one containing the news articles from Positive Data and the other one containing the news articles from Negative Data. Each one of the datasets contains 5000 news Articles.

6. Analyzing the Datasets

For analyzing the Datasets, we have developed a Python application that will execute Term Frequency - Inverse Document Frequency for our two generated Dataset files from the News articles crawled.

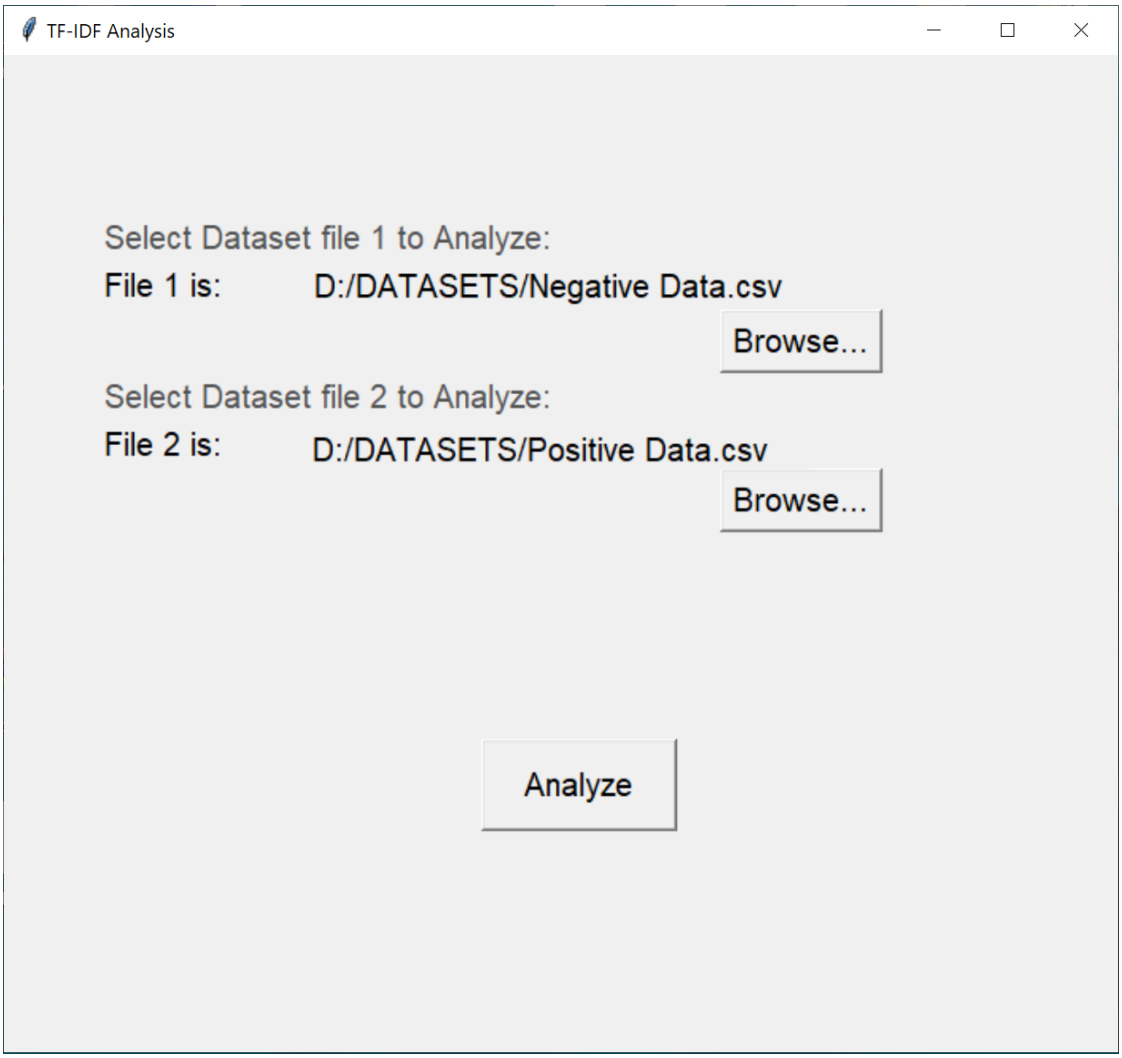

Figure 35. TF-IDF Analysis application

As shown in Figure 35, the interface is very simple to use and offers the options for browsing and selecting two Dataset files.

When the user has selected the files, the interface will show the selected files and their exact location on the computer.

```
def add_file1(self):
        filename = fd.askopenfilename(initialdir = "/",
                                title = "Select file",
                                filetypes = (("csv files", "*.csv"), ("all files", "*.*")))

        self.lbl13.config(text = filename, width=30)
        self.lbl13.update_idletasks()
        self.filepath1 = filename

    def add_file2(self):
        filename = fd.askopenfilename(initialdir = "/",

                                title = "Select file",
                                filetypes = (("csv files", "*.csv"), ("all files", "*.*")))

        self.lbl23.config(text = filename, width=30)
        self.lbl23.update_idletasks()

        self.filepath2 = filename
```

Since our datasets contain useless content in the context of Content analysis, such as “Media Name,” “URL,” and “Date Published,” the application ignores the above-mentioned fields and focuses only on the News titles and News content with their values “text_title” and “text_content” which are text-rich, for getting more relevant results.

```
    def get_text(self, url):
        col_list = ["text_title", "text_content"]
        df = pd.read_csv(url, usecols=col_list)

        titles = df["text_title"].T.values.tolist()
        contents = df["text_content"].T.values.tolist()

        doc_string = ""
        for title in titles:
            doc_string += str(title)
        for content in contents:
            doc_string += str(content)
        return doc_string
```

Since we have to deal with huge amounts of unstructured text data and the TF-IDF results may be unsatisfying, we apply a "Stop words" list, where we add bad and irrelevant words to get more accurate results. The "Stop words" list is created in a separate file from the code file, and in our case, we have it saved as "albanian.txt".

```
        stop_words = []

        for x in open('albanian.txt', 'r', encoding="utf-8").read().split('\n'):
            # print(x)
            stop_words.append(x)
```

We have added to the "Stop words" list irrelevant and often used words such as prepositions, pronouns, phrasal words, question tags, misspelled words, etc.
Below is a short example from around 300 stop words we have added to the list:

'e', 'te', 'i', 'me', 'apo', 're', 'që', 'më', 'gjë', 'një', 'në', 'të', 'keni', 'ke', 'di', 'çdo', 'herë', 'këto', 'siç', 'ja', 'duke', 'ndër', 'kur', 'ketë', 'këtë', 'vitin', 'bën', 'kësaj', 'përmes', 'tuaj', 'tonë', 'madh', 'gjithashtu', 'janë', 'per', 'plote', 'cte' 'sot', 'pamjet', 'ndërsa', 'ish', 'çfarë'. etc

After the user selects the two dataset files and the “Stop words” list is filled, everything is ready. We utilize “TfidfVectorizer” to convert text into Vectors and execute the overall weightage of words from our dataset files.

The “Analyze” button executes the following code:

```
        vectorizer = TfidfVectorizer(analyzer='word', stop_words = stop_words)
        X1 = vectorizer.fit_transform([documentA])
        top_n = 10

        top10_list1 = sorted(list(zip(vectorizer.get_feature_names(), X1.sum(0).getA1())),
key=lambda x: x[1], reverse=True)[:top_n]
        denselist1 = [[i for i, j in top10_list1]]
        feature_names1 = [j for i, j in top10_list1]
        df1 = pd.DataFrame(denselist1, columns=feature_names1)

        X2 = vectorizer.fit_transform([documentB])
        top10_list2 = sorted(list(zip(vectorizer.get_feature_names(), X2.sum(0).getA1())),
key=lambda x: x[1], reverse=True)[:top_n]
        denselist2 = [[i for i, j in top10_list2]]
        feature_names2 = [j for i, j in top10_list2]
        df2 = pd.DataFrame(denselist2, columns=feature_names2)

        print("The result for document 1 is:") print(feature_names1)
        print(denselist1[0])
        print("The result for document 2 is:")
        print(feature_names2)
        print(denselist2[0])
```

For our two datasets, below are the results for the top 10 most used words on both datasets. We can execute both with and without a “Stop words” list, but with vastly different results:

- Document 1 is: Negative Data.csv

- Document 2 is: Positive Data.csv

After executing without a “Stop words” list, we get the following results.

The result for document 1 is:

[0.7465406421378828, 0.3847555617172166, 0.24758846699962217, 0.24484122579517364, 0.15171861103277012, 0.15070833523500518, 0.1327626467220753, 0.11650961327382177, 0.11393075136900074, 0.10638026909096802]

['të', 'në', 'dhe', 'për', 'me', 'që', 'një', 'se', 'nga', 'ka']

The result for document 2 is:

[0.6531016932257355, 0.382904158615431, 0.25647283651835, 0.22016205917507034, 0.20519915462481994, 0.18901296176637358, 0.17609384377981638, 0.17606400747961648, 0.15716271130297912, 0.1292060980156718]

['të', 'në', 'dhe', 'me', 'për', 'që', 'një', 'ka', 'se', 'nga']

Without a “Stop words” list, we notice the poor and unsatisfying results, filled with prepositions, pronouns, phrasal words, question tags, etc., which do not represent real meaning and value to the results generated.
Thus, we continue to execute with a list of "Stop words" for the same two dataset files.

After executing with our custom “Stop words” list, we get the following results:

The result for document 1 is:

[0.16429113130762546, 0.15498374067339335, 0.12214642989643955, 0.12057776855359144, 0.09495629995373897, 0.09087778046233388, 0.08930911911948577, 0.08533517705093722, 0.08376651570808911, 0.08334820601666294]

['shtetit', 'kosovës', '19', 'covid', 'kundër', 'blinken', 'veriut', 'shba', 'zgjedhjeve', 'shqipëria']

The result for document 2 is:

[0.1593912584995167, 0.1283234708258821, 0.11278957698906479, 0.1073864834806066, 0.09607375644727226, 0.08982642957811747, 0.08611180279105246, 0.08391679605324132, 0.0829037160204054, 0.08155294264329085]

['kosovës', 'berisha', 'shqiptare', 'video', 'hapesiren', 'foto', 'pd', 'kurti', 'rama', 'shqiptar']

We can notice that most of the hot topics in the Fake News dataset are Politics related with names of politicians such as 'Berisha,' 'Kurti,' and 'Rama' being some of the most used. If we look inside our Fake News dataset, we can see titles such as:

- *"Zb ulohet id eja nga po licia: Sali Berisha ka pë rgaditur një sk emë të la rgimit nga vendi sikurse Gruevski nga Maqedonia e Ve riut"*
- *"Albin Kurti pyetet de a ka qenë pjesë e U ÇK-së…Pergjigja e tij "tr ondit" te gjith !!!"*
- *"Rama: BE ka m arrë "p eng" Kosovën, e tu rpshme por s'kem çfarë bëjmë"*
- *"Kam të dhëna nga brenda PD", Artur Zheji paralajmëron 'tërmet' në selinë blu*

7. Conclusion and Future work

Although the issue of Fake News was identified many years ago, unfortunately, it is still present as a phenomenon, especially for News in the Albanian language.

The idea for this work was to build a tool that would collect data from various News websites, and for the collected data, an analytical algorithm to be applied.

With the implementation of this work, we have generated two datasets of news Articles, one containing news articles from Fake News sites and the other one containing articles from real Media outlets.

For the two datasets generated, we applied the TF-IDF algorithm to achieve an overall view of the most used terms in the world of Fake and Real News. The results show us that Fake News sites in Albanian are mainly focused on the field of Politics, with some of the main keywords being names of Politicians, whereas in the results from the Real news data, we can mostly see general keywords from various fields.

Although this work has given a general view of the Fake News websites in Albanian, there is a possibility of further advancing our work by crawling data also from social media such as Facebook and Twitter by hashtags.

Also, since a great amount of noise can be noticed in the data, Clustering algorithms can be further used as a pre-processing step for data cleaning.

Bibliography

[1] L. F. S. B. T. S.-E. Thorsten Quandt, "Fake News," *The International Encyclopedia of Journalism Studies,* April 2019.

[2] Hootsuite Inc., "New report finds 1.3 million new users joined social media every day during 2020: 15 new users every second," Intrado GlobeNewswire, Vancouver, 2021.

[3] GCFGlobal.org, "The Now - What is Fake News?," [Online]. Available: https://edu.gcfglobal.org/en/thenow/what-is-fake-news/1/. [Accessed February 2021].

[4] M. DIMOCK, "An update on our research into trust, facts and democracy," Pew Research Center, 2019.

[5] I. L. Emily Dreyfuss, "Facebook Is Changing News Feed (Again) to Stop Fake News," WIRED.com, October 2019. [Online]. Available: https://www.wired.com/story/facebook-click-gap-news-feed-changes/. [Accessed March 2021].

[6] A. Mosseri, "Working to Stop Misinformation and False News," Facebook Newsroom, 2017.

[7] K. Coleman, "Introducing Birdwatch, a community-based approach to misinformation," Twitter, 2021.

[8] B. H. M. M. D. R. D. J. W. Jennifer Allen, "Evaluating the fake news problem at the scale of the information ecosystem," *SCIENCE ADVANCES,* vol. 6, no. 14, 03 April 2020.

[9] C. Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook," 16 November 2016. [Online]. Available: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook. [Accessed March 2021].

[10] I. W.-M. Heather C. Hughes, "The Macedonian Fake News Industry and the 2016 US Election," *PS: Political Science & Politics,* vol. 54, no. 1, August 2020.

[11] R. B. A. K. Salman Bin Naeem, "An exploration of how fake news is taking over social media and putting public health at risk," *Health Information and Libraries Journal,* 12 July 2020.

[12] WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, IFRC, *Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation,* World Health Organization, 2020.

[13] World Health Organization, *Let's flatten the infodemic curve,* World Health Organization, 2020.

[14] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Business Horizons,* vol. 60, no. 3, pp. 293-303, May-June 2017.

[15] I. A. T. H. Z. I. W. K. M. A. M. A. M. S. a. A. G. Ibrar Yaqoob, "Big Data: Survey, Technologies, Opportunities, and Challenges," *The Scientific World Journal,* vol. 2014, July 2014.

[16] A. Emrouznejad, "Setting Up a Big Data Project: Challenges, Opportunities, Technologies and Optimization".

[17] WHISHWORKS, "UNDERSTANDING THE 3 VS OF BIG DATA – VOLUME, VELOCITY AND VARIETY," 08 September 2017. [Online]. Available: https://www.whishworks.com/blog/data-analytics/understanding-the-3-vs-of-big-data-volume-velocity-and-variety/. [Accessed March 2021].

[18] K. Jeffery, "Data Curation and Preservation," in *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences*, vol. 12003, Springer, Cham, 2020, pp. 123-139.

[19] S. Ruggles, "The Importance of Data Curation," in *The Palgrave Handbook of Survey Research*, Palgrave Macmillan, Cham, 2017, pp. 303-308.

[20] ProWebScraper, "What is Data Curation, and Why is it Important?," [Online]. Available: https://prowebscraper.com/blog/what-is-data-curation-and-why-is-it-important/. [Accessed March 2021].

[21] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters,* vol. 31, no. 8, pp. 651-666, 1 June 2010.

[22] A. S. M. C. S. B. Pranav Nerurkara, "Empirical Analysis of Data Clustering Algorithms," *Procedia Computer Science,* vol. 125, pp. 770-779, January 2018.

[23] C. Whittaker, "7 Innovative Uses of Clustering Algorithms in the Real World," April 2019. [Online]. Available: https://datafloq.com/read/7-innovative-uses-of-clustering-algorithms/6224. [Accessed April 2021].

[24] N. A. Z. T. A. A. I. K. A. Y. Z. S. F. A. B. Adil Fahad, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," *IEEE Transactions on Emerging Topics in Computing,* vol. 2, no. 3, pp. 267 - 279, September 2014.

[25] P. K. J. S. A. Z. Hans-Peter Kriegel, "Density-based clustering," *WIREs Data Mining and Knowledge Discovery,* vol. 1, no. 3, pp. 231-240, June 2011.

[26] C. B.-S. Charles Bouveyron, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis,* vol. 71, pp. 52-78, March 2014.

[27] A. M. Kamalpreet Bindra, "A detailed study of clustering algorithms," *6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions),* pp. 371-376, September 2017.

[28] D. M. J. Garbade, "Understanding K-means Clustering in Machine Learning," 12 September 2018. [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. [Accessed April 2021].

[29] I. Dabbura, "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks," 17 September 2018. [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. [Accessed April 2021].

[30] S. Iyyer, "Step by Step KMeans Explained in Detail," April 2019. [Online]. Available: https://www.kaggle.com/shrutimechlearn/step-by-step-kmeans-explained-in-detail. [Accessed April 2021].

[31] E. A. T. M. H. G. Rejab Hajlaoui, "An adjusted K-medoids clustering algorithm for effective stability in vehicular ad hoc networks," *International Journal of Communication Systems,* vol. 32, no. 12, June 2019.

[32] R. N. M. C. Vaggos Chatziafratis, "Hierarchical Clustering with Structural Constraints," *35th International Conference on Machine Learning (ICML 2018),* 2018.

[33] F. Nielsen, "Hierarchical Clustering," in *Introduction to HPC with MPI for Data Science*, Springer, Cham, 2016, pp. 195-211.

[34] Great Learning Team, "What is Hierarchical Clustering? An Introduction to Hierarchical Clustering," 11 July 2020. [Online]. Available: https://www.mygreatlearning.com/blog/hierarchical-clustering/. [Accessed April 2021].

[35] "Understand Jaccard Index, Jaccard Similarity in Minutes," [Online]. Available: https://medium.com/data-science-bootcamp/understand-jaccard-index-jaccard-similarity-in-minutes-25a703fbf9d7.

[36] S. Prabhakaran, "Cosine Similarity – Understanding the math and how it works (with python codes)," [Online]. Available: https://www.machinelearningplus.com/nlp/cosine-similarity/. [Accessed May 2021].

[37] S. V. Apra Mishra, "Analysis of TF-IDF Model and its Variant for Document Retrieval," *International Conference on Computational Intelligence and Communication Networks (CICN),* pp. 772-776, December 2015.

[38] Y.-x. S. Z.-q. W. Y.-Q. Y. Cai-zhi Liu, "Research of Text Classification Based on Improved TF-IDF Algorithm," *IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE),* August 2018.

[39] V. Jayaswal, "Text Vectorization: Term Frequency — Inverse Document Frequency (TFIDF)," October 2020. [Online]. Available: https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d. [Accessed May 2021].

[40] "Beginners Guide to Topic Modeling in Python," August 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/. [Accessed May 2021].

[41] J. Xu, "Topic Modeling with LSA, PLSA, LDA & lda2Vec," May 2018. [Online]. Available: https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05. [Accessed May 2021].

[42] Y. Z. C. W. B. R. Zhijun Chen, "Understanding Individualization Driving States via Latent Dirichlet Allocation Model," *IEEE Intelligent Transportation Systems Magazine,* pp. 41 - 53, March 2019.

[43] H. Z. Zhou Tong, "A Text Mining Research Based on LDA Topic Modelling," May 2016.

[44] F. Revert, "An overview of topics extraction in Python with LDA," December 2018. [Online]. Available: https://towardsdatascience.com/the-complete-guide-for-topics-extraction-in-python-a6aaa6cedbbc. [Accessed May 2021].

[45] V. Choubey, "Topic Modelling Using NMF," July 2020. [Online]. Available: https://medium.com/voice-tech-podcast/topic-modelling-using-nmf-2f510d962b6e. [Accessed May 2021].

[46] R. Chawla, "Topic Modeling with LDA and NMF on the ABC News Headlines dataset," July 2017. [Online]. Available: https://medium.com/ml2vec/topic-modeling-is-an-unsupervised-learning-approach-to-clustering-documents-to-discover-topics-fdfbf30e27df. [Accessed May 2021].

[47] Wordpress.org, "Wordpress," [Online]. Available: https://wordpress.org/about/. [Accessed May 2021].

[48] RSS.com, "How Do RSS Feeds Work?," [Online]. Available: https://rss.com/blog/how-do-rss-feeds-work/.