THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 4. Issue 1 (2024)
https://doi.org/10.62019/abbdm.v4i1.118.


# RSS Feeds Filtering from Multiple Sources Using Automated Techniques of Natural Language Processing

Mujeeb-ur-Rehman Jamali*, Najma Imtiaz Ali, Mujeeb-u-Rehman Maree, Akhtar Hussain Soomro, Safeeullah Soomro, Abdul Ghafoor Memon

## Chronicle

**Article history**

**Received**: February 23, 2024
**Received in t**he **revised format:** March 13, 2024
**Accepted:** March 18, 2024
**Available online:** March 19, 2024

**Mujeeb-ur-Rehman Jamali and Abdul Ghafoor Memon** are currently affiliated with the Emaan Institute of Management and Sciences, Karachi Pakistan.
**Email:** drmujeeb@emaan.edu.pk
**Email:** deancs@emaan.edu.pk
**Najma Imtiaz Ali and Mujeeb-u-Rehman Maree** are currently affiliated with the University of Sindh Jamshoro, Pakistan.
**Email:**najma.channa@usindh.edu.pk
**Email:** mujeeb@usindh.edu.pk
**Akhtar Hussain Soomro** is currently affiliated with the Government College University Hyderabad, Pakistan.
**Email:** akhtar.soomro@gcuh.edu.pk
**Safeeullah Soomro** is currently affiliated with the American National University and University of FairFax USA.
Email: ssoomro@an.edu

***Corresponding Author:**

## Abstract

The Internet's rapid growth has resulted in a surge of information, posing challenges for users trying to stay updated. The World Wide Web, with its diverse websites updated at varying intervals, necessitates efficient categorization of news articles. As the number of online news sources continues to rise, categorization becomes crucial for users to find information easily. Given the overwhelming volume of digital media information, categorizing news using algorithms and assigning multiple tags based on their category becomes essential. In the modern digital era, users face difficulties filtering and accessing relevant content tailored to their preferences due to the abundance of information from various sources. This study employs Natural Language Processing (NLP) methods for automated content filtering of RSS feeds. The developed system extracts feeds via RSS, generates headings and summaries, and employs NLP for categorization. This research has broad implications, offering opportunities for improved information management, personalized recommendations, and informed decision-making across diverse domains.

**Keywords:** RSS Feeds, Content Filtering, Natural Language Processing.



## INTRODUCTION

The internet's rapid expansion has led to an explosion of information. Due of its overwhelming nature, users may find it difficult to remain current on the newest news and information. Social media platforms are a threat to society because they allow inaccurate information, misleading material, and fake news to proliferate. It is harmful to people and is abused in a variety of situations, including cybercrime and political propaganda. The information material is served and organized by a variety of structures, including websites, applications, and information channels. Diverse information is available at unique sites. In order to facilitate user navigation, it could now be useful to

tag the unstructured material. The process of classifying unstructured data involves determining and allocating labels or lessons to it according to its semantic content. This makes it possible to quickly get the chosen data set and to index the data set, which saves time and is enjoyable.

## Automated Content Filtering

Technology-based solutions are used in automated content filtering to automatically filter and categorize content. It is the process of identifying and eliminating unwanted content from a large corpus of text using computer algorithms. In order to analyze and process the data effectively, this approach uses algorithms and techniques. By automating the process, users could find the most relevant information more quickly and with less effort. To do this, regular expression filtering, keyword filtering, and natural language processing techniques are usually combined. Using automatic content filtering, consumers are able to view less undesired or irrelevant content, which makes information management and display easier. By offering tailored information that is in line with consumers' preferences and interests, it seeks to enhance the user experience. Email systems, online forums, social networking sites, news websites, and content management systems are just a few areas where automated content screening may be used.

## RSS FEEDS

Users may subscribe to information from their favorite websites or blogs and receive updates automatically thanks to a technique called RSS (Really Simple Syndication, also known as Rich Site Summary). An RSS feed, which follows a standard format, summarizes the most recent material from a website, including blog entries, podcasts, videos, and news items. It is necessary to use an RSS reader or aggregator, which can be an independent program, a web service, or even a function included in certain email programs or web browsers, in order to subscribe to an RSS feed. The RSS reader aggregates the feeds it has subscribed to and periodically scans them for changes. You don't need to visit every website separately to remain current with them all thanks to RSS feeds. Instead of actively searching for fresh material, the RSS reader automatically pulls the most current changes from the subscribed feeds. This allows a significant quantity of information from several sources to be managed and consumed in one place. RSS feeds are critical to the dissemination and aggregation of material.

Users may subscribe to get information in a standardized manner from several sources. Users can use RSS feeds, which provide a useful way to compile information from several sources, to keep current with their favorite material. RSS feeds are files (XML-based) that contain text and metadata summaries from blogs, news sites, websites, and other online sources. They are set up to provide updates and a standardized framework for material dissemination. By enabling consumers to obtain material from several sources in one convenient area, RSS feeds aim to simplify content aggregation. The RSS Feed Tags are displayed in Table 1. There are several advantages using RSS feeds for content filtering. They offer a uniform format, which simplifies the process of compiling data from several sources. With RSS feeds, users may also customize their content selections and restrict their news feed subscriptions to the subjects that are most relevant to them. Nevertheless, RSS feeds are subject to some limitations. Their primary text-based nature may make them unsuitable for gathering multimedia information, such films or photographs. The reliability

and accessibility of RSS feeds are also impacted by how publishers deploy and maintain them.

**Table 1.**
**RSS Feeds Tags**

| RSS Tag | Description |
|---|---|
| <rss> | Start RSS information |
| <channel> | Channel of the feed publisher |
| <item> | Chunks of summarized information |
| <title> | The title of the information chunk |
| <link> | The url to the html website |
| <description> | The phrases or sentences describing the information. |
| <ttl> | Time to live indicates the amount of time (in minutes) that indicates how long a channel should be cached before refreshing from the source. |
| <pubDate> | The publication date for the content in the channel. |
| <enclosure> | A media-file to be included with an item(Podcast technology). |

XML is used to write RSS feeds. A feed is composed of several articles arranged after a channel. A sample channel looks like the one in Fig. 1:

```
<rss version="2.0">
  <channel>
    <title>…</title>
    <link>…</link>
    <description>…</description>
    <language>…</language>
    <pubDate>…</pubDate>
    <lastBuildDate>…</lastBuildDate>
    <docs>…</docs>
    <generator>…</generator>
    <item>
       <title>…</title>
       <link>…</link>
       <description>…</description>
       <pubDate>…</pubDate>
    </item>
  </channel>
</rss>
```

**Figure 1.**
**RSS Feeds**

## Natural Language Processing

Natural language processing is an artificial intelligence area that focuses on understanding and interpreting human language. Automated text analysis, interpretation, and meaning extraction are all made possible by NLP techniques. NLP techniques can be used in the context of content filtering to glean useful information

from the gathered data and make choices regarding content categorization. The most basic kind of NLP-based content filtering is keyword filtering. Finding keywords that are pertinent to the user's interests and then filtering out content that doesn't contain those keywords. To determine the parts of speech of words in a text, use the part-of-speech tagging technique. A named entity recognition technique can be used to locate named entities in a text, including people, places, and organizations. Based on the entities that are present, this information can then be used to filter content.

## LITERATURE REVIEW

The authors (Yash et al, 2021) proposed a machine learning-based classification method that uses the categories of technology, business, politics, sports, entertainment, and other industries to group the articles. Time is saved and news overhead is decreased by using the news classification system to help readers find interesting news articles. The authors used logistic regression and four-layer artificial neural networks in their suggested method, which is shown in Fig. 2, to predict the class. It uses a continuous bag of words to forecast missing words and skip gram to determine the word's context. Glove embedding is used to create the vector from the words. Without a doubt, each of these recommendations improves the precision of the system. A set of text excerpts selected from various sources serves as the basis for categorizing messages. Prior to being put together in a format suitable for the use of algorithms, the acquired data has to be initialized. Therefore, it is possible to describe each phase of the message classification process, including message collection, initialization, feature extraction, application of the categorization algorithm, and analysis of all performance indicators.

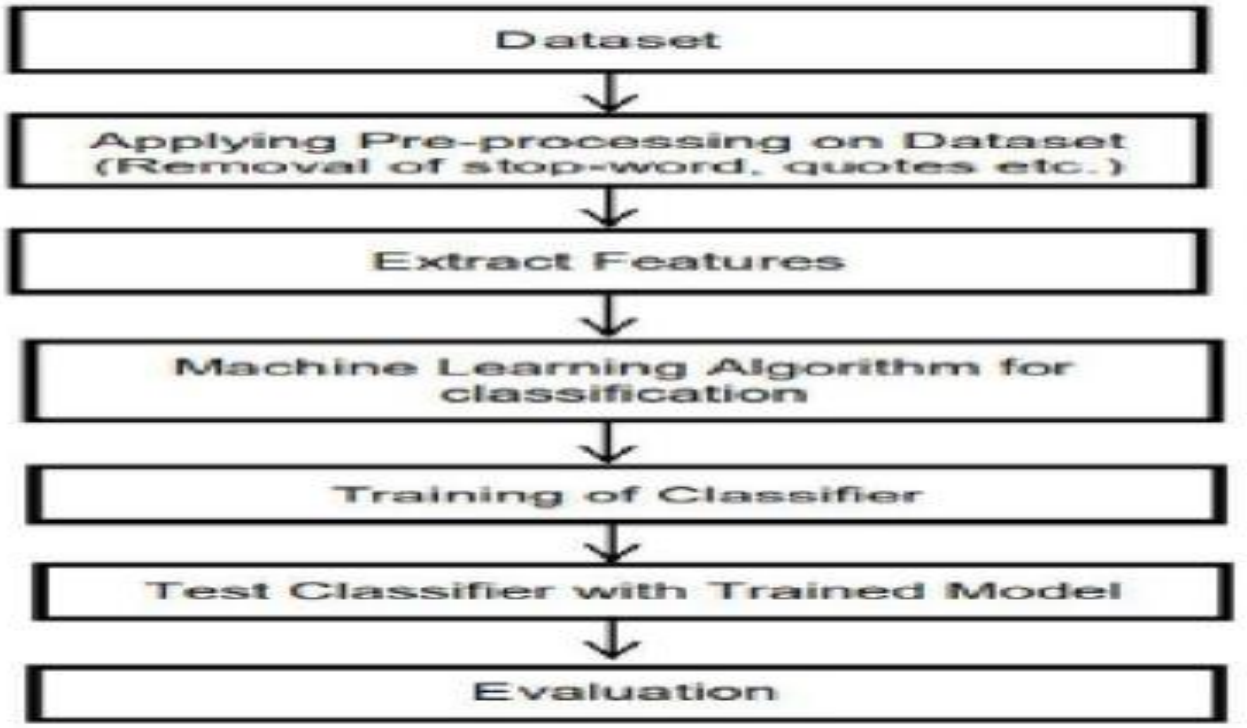


**Figure. 2**
**Text Classification Process**

The authors of (Fong et al, 2010) state that content filtering aims to keep undesirable content from getting to the user. Currently available software content filters typically use an access control list, which requires human search, collection, and classification of unwanted websites before the software filter may prevent access to these URLs. Intelligent categorization and automated web page crawling are the two key components that make up an offline filtering agent. The intelligent classification and automated web page crawling modules are the two key components of Offline Filtering Agent, as shown in Fig 3. Authors proposed a novel method for content extraction from web pages based on an RSS index (Pinheiro et al, 2009). Look through the RSS feed for a list of websites that share a lot of structural similarities, then use the algorithm to get the page template. Determine the body template by calculating the feature of the content block. Ultimately, you will obtain a batch extraction from this set of webpages. The approach offers a high failure tolerance for the Web documents. The results showed that the algorithm is very accurate and widely adaptable.

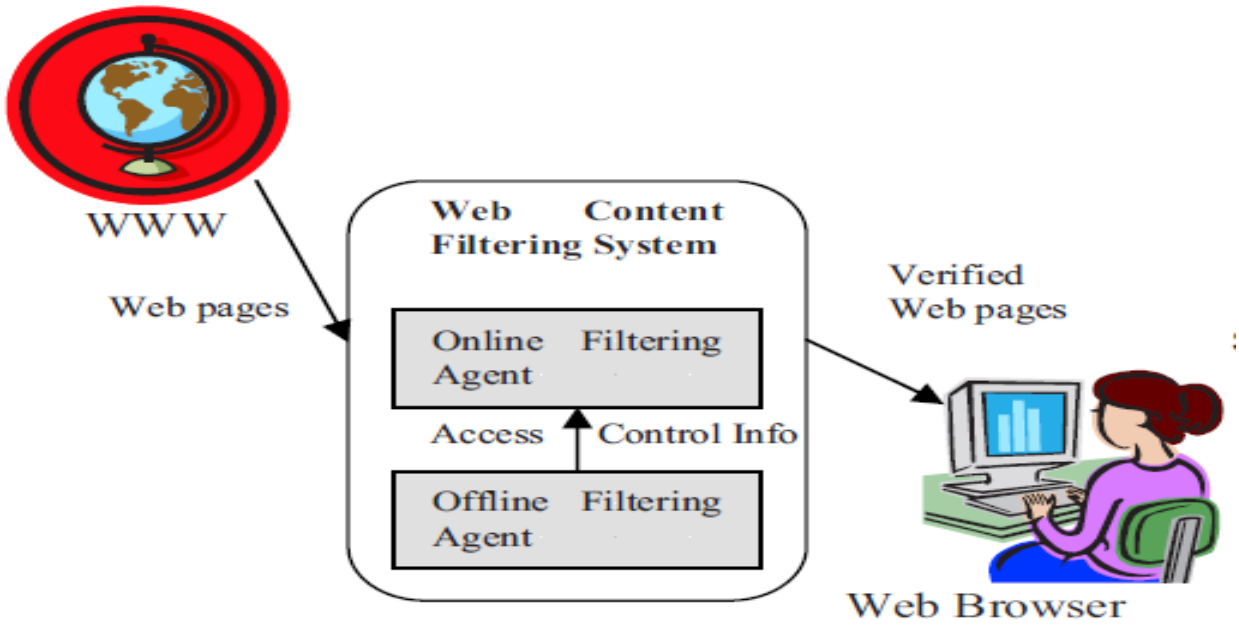


**Figure 3.**
**System for Content and Offline Filtering**

The idea of a page template is combined with the expression "generate once and use repeatedly". It also saves time and effort by eliminating the need to parse HTML texts and correct mistakes. Because of the RSS characteristic, the system is realistic, flexible, and widely applicable. Due to the approximate nature of the procedure used to extract templates from web pages, the efficacy and usefulness of the template are determined by a random selection of samples. Research on how to decrease disparity, improve algorithm efficiency, align generated templates more closely with real templates, and pinpoint the body content more precisely will be conducted in the future. Online page filtering and recommender systems both frequently utilize content-based filtering. The fundamental idea of content-based filtering (Adomavicius and G. Tuzhilin, 2005). The writers provided an overview of recommender systems and discussed the most recent generation of recommendation strategies, which may be broadly divided into three groups: collaborative, content-based, and hybrid recommendation approaches. The

authors of (Hui et al, 2012) proposed a unique way for recommendation system using Hidden Markov Model by combining a probabilistic model with conventional content-based filtering recommendation algorithms. This technique's main idea is to determine how well an object profile matches a user profile, then suggest items that best suit the user's needs or preferences (Chmielewski and Hu, 2005).

## MATERIAL AND METHODS

Efforts were made to conduct research and create a reliable system for gathering and handling RSS feeds. To aggregate and process RSS feeds from various sources, a system is designed. The system to be able to gather and compile information from various websites and blogs into a single, cohesive format. to investigate and use NLP methods to examine the content that has been gathered. To extract valuable information and insights from the retrieved articles or posts, this includes techniques like text classification, sentiment analysis, topic modeling, and entity recognition. Create a content filtering algorithm with intelligence that classifies and ranks the content according to user preferences, relevance, and quality using the results of the NLP analysis. In order to continuously increase the filtering accuracy over time, the algorithm dynamically adapt to user feedback. to create a user-friendly interface that makes it easy for users to browse and consume the filtered content. The interface allowed for easy navigation and interaction with the content, as well as personalized recommendations and preferences customization for users.

The proposed system has the potential to improve information retrieval effectiveness by allowing users to access pertinent and trustworthy content from numerous sources without feeling overloaded by the vast amount of information available. Using natural language processing (NLP) and RSS feeds, the project aims to address the challenge of creating an automated content filtering system that can filter material from various sources. Utilizing RSS feeds and natural language processing methods, the approach for automating content filtering from many sources locates the most recent studies on the subject and performs a literature review. This made it easier to determine the automated content filtering state of the art at the time and the knowledge gaps that required further research. Gathering of information for testing an automated content filtering system. This information to be gathered from news articles, RSS feeds, or other sources to extract features from the data for the automated content filtering system to be trained which depending on the text's content. Utilizing the extracted features, train the automated content filtering system. Using the information gathered, evaluate the automated content filtering system.

By evaluating the system's precision, recall, and accuracy, this can be achieved. the research's findings to be interpreted and the future research directions to be determined. The methodology for this research work will use NLP techniques. The methods include text classification, sentiment analysis, topic modeling, entity recognition, or any other NLP approaches appropriate for the content filtering task. The study's overall methodology will be followed, as specified by the research design. It entails choosing the research techniques, data sources, and overall structure for putting the automated content filtering system into practice. In this study, research techniques are used to automatically filter content from various sources using RSS feeds and NLP techniques. The research

approach involves running tests to evaluate the performance of various automated content-filtering strategies. Both the lab and the real world can be used for experiments. They are surveyed about automated content filtering strategies to get their opinions. Surveys methods use to learn what users want from automated content filtering and how they view the efficacy of various methods. The case studies approach involves examining how automated content filtering is applied in actual scenarios. Case studies can be used to pinpoint the difficulties and achievements of putting automated content filtering into practice in various contexts. Reviewing the body of knowledge on automated content filtering is part of a literature review.

A literature review to be used to determine the automated content filtering state of the art and to identify any knowledge gaps that require further study. The need for large datasets was one issue that needed to be resolved in order to research automated content filtering. Large datasets of RSS feeds and news articles are evaluated using automated content filtering methods. In order to automate content filtering, it is necessary to extract the meaning of text using NLP techniques. The methods and tools that are used for automated content filtering from numerous sources using RSS feeds and natural language processing techniques include sentiment analysis, named entity recognition, part-of-speech tagging, keyword filtering, and part-of-speech tagging.

### RSS Feed Extraction

The system uses libraries to extract feeds from various sources. This first stage is obtaining up-to-date content from numerous websites or sources that publish information via RSS. Headline and Summary Extraction: The retrieved RSS feed is parsed to obtain headlines and summaries for each article or item of material. This is critical for further NLP analysis since it contains textual data for content classification. Text Classification with Machine Learning: Machine learning-based text categorization algorithms are used to divide the material into predetermined groups. These algorithms are trained on labeled data to identify patterns and relationships, allowing for reliable categorization.

A research design combines relevance with the objective of the study by organizing conditions, gathering data, and analyzing it (Jamali, M.-u.-R. et al, 2021). The conceptual framework that a study is performed inside is known as its research design. In particular, the kind and type of data needed, the sources from which the necessary data may be obtained, data collection methodologies, data analysis techniques, and report preparation were all incorporated in the research design. The study procedures and methodologies used are shown in Fig. 4 (Jamali, M.-u.-R. et al, 2024), (Jamali, M. R. et al, 2020):

## SYSTEM DEVELOPMENT ENVIRONMENT

- Programming Language: Choosing an appropriate programming language that offers robust support for data processing and NLP libraries, such as PHP or Java. The tool dynamically generates a PHP file with rephrased news and heading. A web server is then used to open the PHP file.
- Libraries and Frameworks: The use of NLP frameworks and libraries for feature extraction and data analysis.

- The gathered data, user preferences, and filtered content are stored and managed using the NoSQL database management system i.e., MongoDB.
- The system's interactive user interface is built using the User Interface Framework.

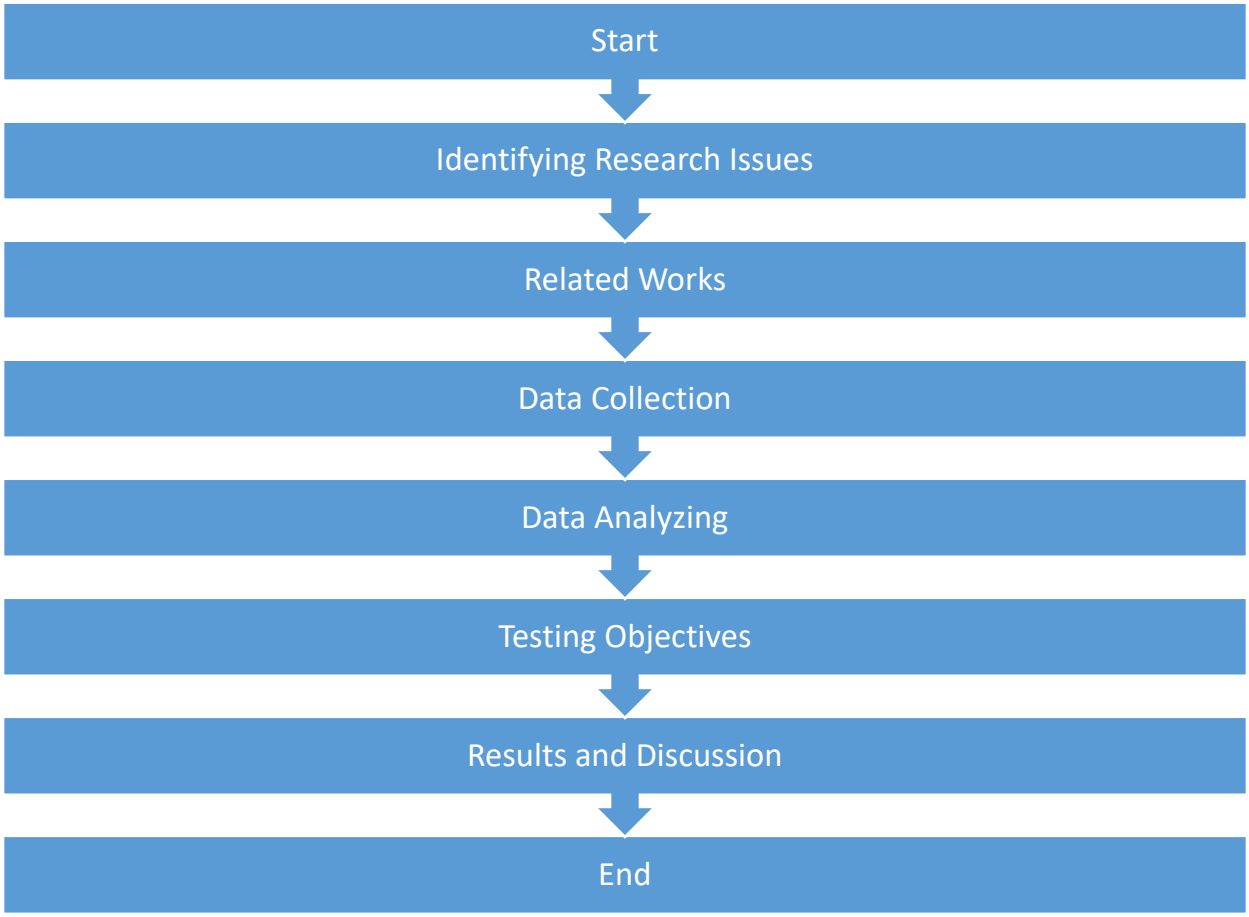


**Figure 4.**
**Techniques and procedures for research**

## PROBLEMS AND MAIN CHALLENGES

- RSS feeds may be overwhelming, making it difficult for consumers to identify relevant and interesting material.
- Diverse Sources: RSS feeds from several sources cover diverse subjects and opinions. Filtering material from many sources need a careful methodology to guarantee complete coverage and relevancy.
- Language Variations: The dynamic nature of language, including variations, subtleties, and changing idioms, makes it challenging to categorize and analyze material.
- User Preferences: Filtering material based on individual preferences can be challenging due to various interests and relevant criteria.
- Timeliness: Efficient algorithms are necessary for processing current information fast.

# SYSTEM DESIGN

Information retrieval, social media filtering, and news aggregation are just a few of the potential uses for the proposed approach. Using RSS feeds and NLP methods, the main goal of this study is to create an automated content filtering system using RSS feeds, system is implemented to gather and process data from various sources. To make sense of the gathered data, it applies NLP techniques. To categorize the content based on user preferences, content filtering algorithms is developed. Analyze the system's functionality and effectiveness in filtering and classifying content. In order to address these issues, this dissertation proposed an automated content filtering system that makes use of RSS feeds and NLP methods. The developed system can quickly and effectively extract and filter pertinent content from various sources, which saves time and effort when retrieving information.

Comparing the proposed method to current ones for automated content filtering, there are a number of benefits. It can first filter content from various sources. The second capability is the ability to extract the meaning of the text using natural language processing methods, enabling more complex filtering. The strategy is scale able to big datasets of RSS feeds and news articles. This developed system to pull the news feed via RSS feeds using the pytorch library and then writing a heading for it and summary for it. The extracted news feed is then fed to an NLP system. Each news is assigned a heading for that particular day. All the news (summarized and heading) are stored in a file system locally Probably readable only by the User Interface (UI). The system showing the heading and probably details of news providers - fetched from the file system. User click the heading thereafter summarized news in another window fetched from the file system.

# SYSTEM DESCRIPTION

A content filtering system's system architecture includes the following elements:

**Data collection:** The data needed to perform the filtering to be gathered by this component.

**Data preprocessing:** This section is in charge of clearing the data of any duplicate or irrelevant information.

**Feature extraction:** In this part, features are taken out of the data.

**Classification:** The data to be categorized by this component using a machine learning algorithm.

**Decision-making:** This element is in charge of choosing what to do with the unwanted content.

**User interface:** Users can communicate with this element and receive feedback about the content filtering procedure. The dispersed platform that allows web masters (administrators) to set up websites to integrate RSS feeds from other websites and users to automatically obtain current information. It is made up of a number of servers, each of which could provide information that has been archived from RSS feeds that it has compiled for usage by other servers. The administrator and user components of the distributed platform are as follows:

- It provides a multitude of choices for the server administrator to personalize every facet of the server's operation, including listening ports, update intervals, and the addition or deletion of additional servers and locally cached feeds.

- For the server administrator, it offers a variety of options to customize every aspect of the server's behavior, such as listening ports, update intervals, and the addition or removal of other servers and locally cached feeds.

- Users can access the data that the server has stored by using a portion of the platform. This is accomplished by first connecting to the server using HTTP and a browser to get a list of the feeds that are accessible both locally and remotely, and then connecting to these feeds using an RSS reader using the knowledge gained through the browser. Individual users are also tracked by the distributed platform. A user can choose whether or not they want the server to keep track of which items have been retrieved when they access a feed. Using the same account, the user will also be able to access data from multiple locations. A group of parts referred to as managers make up the server. With minimal assistance from the others, each of these managers is in charge of carrying out a task, resulting in a modular structure that is simple to build upon.

- Database Manager: In charge of overseeing all interactions with the embedded database. A large number of concurrent threads to be accessed the database while still guaranteeing consistency. Only one connection is permitted to the database in order to achieve objective.

- The port manager manages all socket communications that the server must do. Two types of ports are available. The first is a server socket that waits for requests from other servers. The port manager launches a thread in response to a request, processes it, and replies with the requested information. Outgoing ports are the second kind; they are used to send queries from one server to another to find out whether feeds are accessible, if the server is active, or if the server has any archive material for a channel stored on it.

- The servlet manager controls clients and their requests via the integrated web server. When the program is launched, the servlet manager launches the server. The manager must speak with the database manager in order to obtain feed data and accessible channels. Apache Tomcat is used as an embedded HTTP server.

- Configuration Manager stores data in a file, retrieves it as needed, and allows other managers to access its configuration data. An administrator performs system changes unrelated to database modifications, the configuration manager is notified and the settings in the locally stored file are updated.

- Synchronization Manager regulates when server events occur.

Two main groups can be drawn from these incidents.

Using locally cached feeds for synchronization is the first choice which is required to get the remote website, periodically obtain its feed information, and then input the information into the database. The second method allows different servers to have different synchronization intervals: remote server synchronization. The management has to set a separate timer for each server that want to update often. Interaction with clients:

The client logs on to the platform using any web browser. After logging in to access feeds, the user can select one of the following three options:

- Track which feed items have been retrieved: By selecting this option, the user can update the server's database with the feed items they have retrieved.
- Obtain every item that has become accessible since the previous tracked extraction from the feed. When the user selects this option, all feed data collected since the previous usage is retrieved. When seeing any new items, the client accesses the server from any place; they do not wish to save the feed data.
- This feature enables the user to get every feed item that is currently available.

## DATA COLLECTION AND PREPROCESSING

The selection of pertinent RSS feeds from various sources is a necessary step in the data collection process. Establishing a system to automatically collect RSS feeds and identifying the sources of the desired content are crucial. Once the RSS feeds have been gathered, the next step in data preprocessing is to clean and format the extracted text in order to get it ready for further analysis. In this step, user might also tokenize, handle special characters, remove HTML tags, remove stop words, and so on. Using RSS feeds, the data collection module retrieves content from various sources. Using the user's preferences as a guide, it selects the pertinent feeds and downloads the most recent updates from those sources.

Preprocessing, or cleaning up the data to remove noise and unimportant information, is the process that the collected data goes through. In order to ensure consistency in terms of format and structure, it also entails normalizing the data. Utilizing the specified evaluation metrics, the evaluation results show how well the implemented system is performing. These outcomes reveal how well the system classifies and filters content based on user preferences. The outcomes to be displayed as tables or graphs that show the metrics the system was able to achieve. To give a thorough understanding of the system's performance, qualitative analysis can also be added. The performance of the system is covered in the results discussion. It evaluates the outcomes obtained to determine how well the system categorizes and filters content according to user preferences. The evaluation's findings highlight the areas where the system performs well and those where improvement is needed.

They also look at the system's strengths and weaknesses. The system's abilities to precisely identify pertinent content, efficiently summarize data, and adapt to user preferences were among its strong points. On the other hand, there are times when the system is unable to handle a particular type of content. It investigates the causes of particular performance outcomes, such as the effects of particular NLP methods or the effects of dataset characteristics. Performance of the system concentrating on evaluation metrics. It offers numerical evaluations of how well the system does at accurately classifying and filtering content in accordance with user preferences. The overall effectiveness and performance of the system are demonstrated by how the achieved metrics were interpreted. The NLP techniques used and how they affect the system's functionality are important aspects of the system. It looks at how well methods like topic modeling, sentiment analysis, and keyword extraction perform when it comes to sifting out

important features from the content. It also takes into account any modifications or additional methods that might strengthen the feature extraction procedure. The automated content filtering system has restrictions on the dataset used, the NLP techniques chosen, and the effectiveness of the content filtering algorithm. These limitations reveal areas where future work can be improved or expanded upon. The system's potential advantages in content filtering from various sources, personalized recommendations, and information retrieval. The creation of a strong system architecture, the use of NLP for feature extraction, and the enhancement of user experience via the user interface are the contributions made by this research.

With the help of the PyTorch library, system to pull feeds via RSS feeds and a headline and summary for each feed. Next, an NLP system receives the extracted feed. Each item is given the appropriate header for the day. A local file system contains all of the (summarized and headed). likely only accessible through the User Interface (UI). The system displays the headline and likely the details after retrieving them from the file system. User clicks the heading, and then a window displaying summarized that has been fetched from the system. Multilingual Challenges, resource Intensiveness and Adaptability, and Continuous Learning are the system limits. NLP models trained on a single language may not perform as well when dealing with content in numerous languages in Multilingual Challenges. It may be difficult to translate and interpret material in languages other than the model's primary training language. The Consumption of Resources Implementing complex NLP models for real-time filtering of many RSS feeds from various sources can be computationally demanding and resource-intensive.

Adaptability and Continuous Learning: NLP models may struggle to adjust to shifting trends, and continuous learning to update the model in real-time might be difficult. Precision is a quantitative system assessment statistic that is the ratio of relevant instances (properly filtered articles) to total instances anticipated as relevant by the system. High accuracy suggests that the algorithm is likely to be correct when it forecasts as relevant. When evaluating the performance of a system for filtering RSS feeds from multiple sources using automated Natural Language Processing (NLP) techniques, several key evaluation metrics can be defined to assess the system's effectiveness and efficiency, namely execution time, precision, and accuracy. These metrics help to assess many elements of the system's performance.

## CONCLUSION

This research's key findings, which led to the development of an automated content filtering system that efficiently gathers, preprocesses, and filters content from various sources, contribute to the performance of the system as a whole. In addition to using NLP for feature extraction and content categorization, the system uses RSS feeds to retrieve data. Results analysis shows that the system successfully achieves high levels of accuracy when accurately filtering content based on user preferences. This study contributes in a number of ways to the field of automated content filtering. First, suggested an architecture for the system that combines algorithms for data collection, preprocessing, feature extraction, and content filtering. An effective framework for managing content from various sources and personalizing the filtering process is provided by system architecture. To extract valuable features from the content, the system also employs a variety of NLP techniques, including topic modeling, sentiment analysis, and keyword

extraction. These methods improve the process of content filtering's accuracy and relevance. Last but not least, the development system offers a user interface that enables users to interact with the system, specify their preferences, and view the filtered content, thereby improving the user experience.

## DECLARATIONS

**Acknowledgement:** We appreciate the generous support from all the supervisors and their different affiliations.

**Funding:** No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

**Availability of data and material:** In the approach, the data sources for the variables are stated.

**Authors' contributions:** Each author participated equally to the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

**Consent to Participate:** Yes

**Consent for publication and Ethical approval:** Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

## REFERENCES

A. Adomavicius and G. Tuzhilin, (2005). "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749.

A.C.M. Fong, S.C. Hui and G.Y. Hong, (2010), "An intelligent offline filtering agent for website analysis and content rating", pp-1-4.

D. Chmielewski, Gongzhu Hu. "A distributed platform for archiving and retrieving RSS feeds", Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05), 2005.

David Chmielewski and Gongzhu Hu, (2005). "A Distributed Platform for Archiving and Retrieving RSS Feeds", Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05) from IEEE Xplore, pp-1-6.

Hung-Wei Chen, Yi-Leh Wu, Maw-Kae Hor, Cheng-Yuan Tang, (2017). Fully Content-Based Movie Recommender System with Feature Extraction Using Neural Network. Proceedings of the 2017 International Conference on Machine Learning and Cybernetics, Ningbo, China, 9 - 12 July 2017. 978-1-5386-0408-3/17.

Jun Shi Bo Long Liang Zhang-Bee-Chung Chen Deepak AgarwalWeiwei Guo, Huiji Gao. (2019). "Deep natural language processing for search and recommender systems.

Ladda Preechaveerakul and Wichuta Kaewnopparat, (2009). "A Novel Approach: Secure Information Notifying System using RSS Technology", International Conference on Future Networks, from IEEE Xplore, pp-1-5. DOI 10.1109/ICFN.2009.35.

Li, Hui, Fei Cai, and Zhifang Liao, (2012). "Content-Based Filtering Recommendation Algorithm Using HMM", Computational and Information Sciences (ICCIS), Fourth International Conference on. IEEE.

M. R. Jamal, A. G. Memon, M. R. Maree, (2020). "Security issues in data at rest in a non-relational Document Database", Sindh University Research Journal (Science Series), Vol. 52 (03) p-279-284. Http://doi.org/10.26692/sujo/2020.09.41.

Mujeeb-ur-Rehman Jamali, Abdul Ghafoor Memon, Nadeem A. Kanasro, Mujeeb-u-Rehman Maree. (2021). "Data integrity issues and challenges in next generation non-relational document-oriented database outsourced in public cloud", International Journal of Emerging Trends in Engineering Research, Volume 9. No. 4, pp-416-420.

Mujeeb-ur-Rehman Jamali, Najma Imtiaz Ali, Abdul Ghafoor Memon, Mujeeb-u-Rehman Maree and Aadil Jamali (2024). Architectural Design for Data Security in Cloud-based Big Data Systems. Baghdad Sci.J [Internet]. [cited 2024 Mar. 12];21(9). https://bsj.uobaghdad.edu.iq/index.php/BSJ/article/view/8722. DOI:https://doi.org/10.21123/bsj.2024.8722

Mtuthuko Mngomezulu and Ritesh Ajoodha, (2022). "A Content-Based Collaborative Filtering Movie Recommendation System using Keywords Extractions", Proc. of the 8th International Conference on Engineering and Emerging Technologies (ICEET), pp-1-6.

Thushara, M. G., Tadi Mownika, and Ritika Mangamuru, (2019). "A comparative study on different keyword extraction algorithms." 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE.

Thushara MG, Krishnapriya MS, and Sangeetha S Nair, (2017). "Domain classification and tagging of research papers using hybrid key phrase extraction method".

Thushara, M. G., M. S. Krishnapriya, and Sangeetha S. Nair, (2017). "A model for auto-tagging of research papers based on key phrase extraction methods." International conference on advances in computing, communications and informatics (ICACCI). IEEE.

Wallace A. Pinheiro, Thiago de S. Rodrigues1, Marcelo A. R. da Silva1, Márcio A. N. da Silva1, Marcelino Campos Oliveira Silva1, Geraldo Xexéo, Jano M. de Souza, (2009). "Autonomic RSS: Discarding Irrelevant News, Fifth International Conference on Autonomic and Autonomous Systems, pp-148-153. IEEE Xplore. DOI 10.1109/ICAS.2009.11.

Yash Veer Singh, Piyush Naithani, Parvez Ansari, Pragya Agnihotri. "News Classification System using Machine Learning Approach", 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021.