International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Feature Selection Algorithm-Based Data Filtration Model For "Data Journalism"

Syed Irfan Yaqoob*[1], Arun Prakash[2], Anuradha Kanade[3], Shantanu Kanade[4], Neha Bharti[5]

Submitted: 11/01/2024 **Revised**: 17/02/2024 **Accepted**: 25/02/2024

Abstract: Journalism today has become a more dynamic and technology-oriented profession unlike conventional journalism. At the same time, it has also become a challenging task to handle and filter the huge amount of multimedia data received by media houses. It requires a larger workforce to manage and filter the data in order to make good news stories/packages. The media houses are now branching out to multiple platforms. The consumption of news is moving more towards the digital domain, people have also shifted their preference to consuming short, precise and relevant news in a personalized manner. Management of abundance Multimedia data in the media houses can be done in a concise amount of time (FSA) Feature Selection Algorithm is utilized to make the idea of data journalism more effective and efficient, we are offering a model that employs "FSA" to filter the requested (relevant) data from the enormous amount of in-house data.

Keywords: *Data Journalism, Machine Learning, Feature Selection, Data filtration.*

1.Introduction

Data-driven journalism is the conventional name for Data journalism now, it is a new concept in journalism that is used for analyzing the relevance of any news story based on various sets of data (Hintz et al., 2022). Despite growing data volumes and constantly improving natural language processing methods, work on automated journalism is still very rare (Nanekar et al., 2023). Data journalism is an emerging branch of journalism that utilizes a collection of disparate data sources to generate and disseminate information via the mass media. (Li, 2022). Data journalism studies frequently places technology at the core of its techniques by embracing computer programming culture and the rationality of computation (Hintz et al., 2022).

In the era of information and technology every media house receives multimedia data in bulk. (Vijay et al., 2023) There are also various social media platforms that generate an abundance of newsworthy stories and data, Media houses need to keep an eye on them as well so that they do not miss any trending or latest information (Vijay et al., 2023).With the accessibility of smartphones to almost every teenager and adult, the concept of citizen journalism has really taken off resulting in generation of local and hyper local multimedia data (Dai et al., 2022).

Media Organization spends millions of dollars each year on human resources and other infrastructure for processing documents and other forms of multimedia resources manually, to produce relevant and extracted news to public. This Process is time consuming, error prone, expensive and not easy to scale (Rahman & Bockarie, 2022)

"Feature selection machine learning technique" can help in making this expensive and exhaustive task easier, improves data accuracy, and process data faster. We are proposing a model in this paper to automate this system with the specific data driven methodology for making Data Journalism more efficient, effective, increase customer satisfaction, employee productivity and lower cost with intelligent document and other multimedia processing by the help of machine learning and AI technology which will automatically processes and analyses data from different multimedia resources.

With the growing demand for more personalized information, extracting personalized data/ news seems an attractive package to attract more audience/ customer for the news organizations or companies (Morini, 2023). The delivery of personalized news is not merely limited to established news portals or apps; it is also part of the dashboard of technological giants such as Google, Microsoft, Facebook and Amazon etc. (Wu et al., 2023). The delivery of news content is also adopted by Virtual Voice assistants like Alexa and Siri, therefore the mechanism to extract the personalized news can be helpful for the corporate as well. At the same time, the attention time span of the users has also decreased significantly, now there is a demand for short, precise, crisp and relevant news only (Nag et al., 2023).

[1,2,3] *Dr.Vishwanath Karad MIT World Peace University Pune INDIA.*
[4]*Symbiosis School For Online And Digital Learning (Symbiosis International University Pune INDIA*
[5]*Banaras Hindu University Varanasi INDIA*

When it comes to machine learning, feature selection is often used as a part of the preprocessing process. By selecting a part of the original features, you can narrow down the number of features that need to be reduced based on a certain test. (Singh et al., 2024). Since the 1970s, the field of research and development surrounding feature selection has flourished. It has been demonstrated to be successful in eliminating redundant and superfluous data, thereby increasing the efficiency of learning tasks. It has been demonstrated to be successful in eliminating redundant and superfluous data, thereby increasing the efficiency of learning tasks. (Singh et al., 2024). The filter model can be divided into two main categories: the wrapper model and the feature selection algorithms. The filter model can be divided into two main categories: the wrapper model and the feature selection algorithms.The Filter model selects some features without using a learning algorithm by using generic properties of the training data (Feng & Wang, 2023). The wrapper model uses a single pre-defined learning algorithm to evaluate and select features to be used. The wrapper model uses a single pre-defined learning algorithm to evaluate and select features to be used. (Tran, 2019). Recently, hybrid algorithms have been proposed to combine the benefits of both models in order to process large amounts of data. Recently, hybrid algorithms have been proposed to combine the benefits of both models in order to process large amounts of data. In these techniques, the best feature subsets for a specific cardinality are first selected using a goodness measure of feature subsets based on data properties, and then cross validation is used to select a final best subset across various cardinalities (Tran, 2019). The basic goal of these algorithms is to combine filter and wrapper algorithms to get the optimum performance from a given learning algorithm with a time complexity similar to filter algorithms (Lan, 2017).

2.Review of Literature

Despite a thorough assessment of the literature, not much has been done to filter data for data journalism using machine learning. Even though there are some connected research papers and book chapters.

According to Paul Bradshaw in his book chapter discusses about Rock stars can be made with data journalism. Data journalism can entail a variety of people and skills, from those who are strong at gathering information to those who can extract stories from it, write the stories, visualize the information, or create interactive data. Data-driven journalism increasingly involves the development of user-friendly products, including searchable databases and applications. The development of user-friendly products, including searchable databases and applications, in addition to just presenting stories. Utilizing the Freedom of Information Act of 2000 to ask public entities for material, including data and documents, is a terrific approach to obtain information for exclusive articles (Paul Bradshaw, 2017).

Mary Lynn et. al. (2018) The research argues for the excellence of the recipients and nominees in major domestic and global data journalism competitions. Since the 2012 award of the inaugural Data Journalism Prize, a content analysis has been conducted of data projects submitted by Canadian media outlets to three journalistic organizations, namely the Online News Association and the Global Editors Network. The research argues for the excellence of the recipients and nominees in major domestic and global data journalism competitions. Since the 2012 award of the inaugural Data Journalism Prize, a content analysis has been conducted of data projects submitted by Canadian media outlets to three journalistic organizations, namely the Online News Association and the Global Editors Network, and the Canadian Association of Journalists—has been completed. This study examines how journalists developed what might be regarded as outstanding data journalism. Their findings indicate the absence of widely acknowledged criteria for what defines excellence. The use of free online tools like Google Maps had an impact on the projects' quality, which were challenging to tailor to specific needs, and the fact that most practitioners worked within traditional journalism frameworks. Dynamic maps, graphs, and video were the most frequently used visual components. All but one of the projects had an interactive component in terms of interactivity. Within the realm of information visualization, the prevalent interaction methods encompassed inspection and filtering, considered fundamental techniques. These techniques demonstrate the need for collaborative, interdisciplinary research in data journalism and a greater focus on the implications of Google Maps, as well as other practical tools. These techniques demonstrate the need for collaborative, interdisciplinary research in data journalism and a greater focus on the implications of Google Maps, as well as other practical tools..

Groveet al. (2020) in their article have argued that in recent years, academics have investigated how supervised machine learning (SML) may be applied to journalism studies. The discipline may benefit from such computational tools, however the justification for using these supervised models contains several presumptions that require additional investigation. This essay aims to outline the circumstances in which SML might be helpful for academics studying journalism as well as the field's progress in utilising its potential advantages. We begin with an outline of SML's uses in journalism studies before introducing it. We then list the difficulties the field will face in implementing these strategies. This includes over-estimating the amount of time and resources saved by automated coding, and not using appropriate sample

methods. This includes over-estimating the amount of time and resources saved by automated coding, and not using appropriate sample methods. being aware of the risk posed by algorithmic determinism, and having restricted predictive modelling generalizability (Grove, et. al., 2020).

Jacobi et al. (2016) in their paper argued thatthe enormous collections of news items that have become accessible thanks to digital technology both justify and permit scientific investigation, posing a challenge to journalism scholars to analyse previously unheard-of volumes of texts. To meet this difficulty, we suggest Latent Dirichlet Allocation (LDA) topic modelling. Modern content analysis methods like LDA, which track patterns of word (co-)occurrence to identify latent subjects, may automatically arrange huge archives of texts. We describe the operation of this technique, the impact of the researcher's decisions on the outcomes, and the interpretation of the outcomes. In order to demonstrate the usefulness of a case study for journalistic research, we conducted a case study on the nuclear technology coverage of The New York Times from 1945 through the present day, partially reproducing a study. In order to demonstrate the usefulness of a case study for journalistic research, we conducted a case study on the nuclear technology coverage of The New York Times from 1945 through the present day, partially reproducing a study. by Gamson (Jacobi et al. 2016).

Albert et al. (2014) in their article said To gain insights into how journalistic bias relates to the content of an article and how reader demographics and political leanings impact the perceived objectivity of articles, we conducted a study. We used supervised machine learning techniques, including L2-regularized L2-loss SVM, Naive Bayes, and Linear Regression, on data gathered from readers on the Mechanical Turk crowdsourcing platform. This allowed us to predict the average polarity of 284 news articles and how each reader perceived the polarity of a specific article based on both its content and the reader's individual characteristics. (Albert et al., 2014).

Bhoite, S et al. (2023) Says Feature selection in RNNs and LSTMs is not as straightforward as in traditional machine learning models, and it often relies on the model's ability to learn relevant features from the data. Therefore, a significant part of the process involves experimentation and fine-tuning to optimise model performance for your specific task.

Shubham khandale et al. (2019) The selection of a feature selection method is contingent upon the complexity of the problem, the data set, and the machine learning algorithm employed. This process typically necessitates trial and error to identify the most suitable method for selecting the features. The selection of a feature selection method is contingent upon the complexity of the problem, the data set, and the machine learning algorithm employed. This process typically necessitates trial and error to identify the most suitable method for selecting the features.

3. Feature Selection

The process of feature selection involves the selection of the most relevant features from the original set of features, while eliminating any duplicate, redundant, or ambiguous features. Machine learning models are typically constructed using a limited number of variables from a dataset. As others may be redundant or unimportant. Including these redundant or irrelevant features in the dataset can negatively impact the model's overall performance and accuracy. Therefore, It is necessary to select the most appropriate elements from the data by using feature selection within machine learning to eliminate redundant or less relevant elements. (Jawapoint, 2023)

Finding the ideal set of features to enable the development of useful models for researching the phenomenon in question is the objective of machine learning feature selection techniques is to identify the following categories: can be used to classify these techniques:

3.1. Supervised methods: These methods are used on labeled data with the goal of finding pertinent features that improve the functionality of supervised models like regression and classification.

3.2. Unsupervised Methods: These methods are suitable for unlabeled data and are designed to select meaningful features in situations where labeling is not available.These methods fall under the following taxonomic categories:

A. Filter methods

B. Wrapper methods

C. Embedded methods

D. Hybrid methods

Filter Method:

The focus of filter approaches is on the intrinsic properties of features evaluated using univariate statistics, rather than cross validation performance. Filter methods are faster and more efficient than wrapper methods. Additionally, filter methods are more efficient when working with large-scale data. By evaluating the information gain of each variable relative to the desired variable, information gain is used to identify the decrease in entropy caused by data changes.

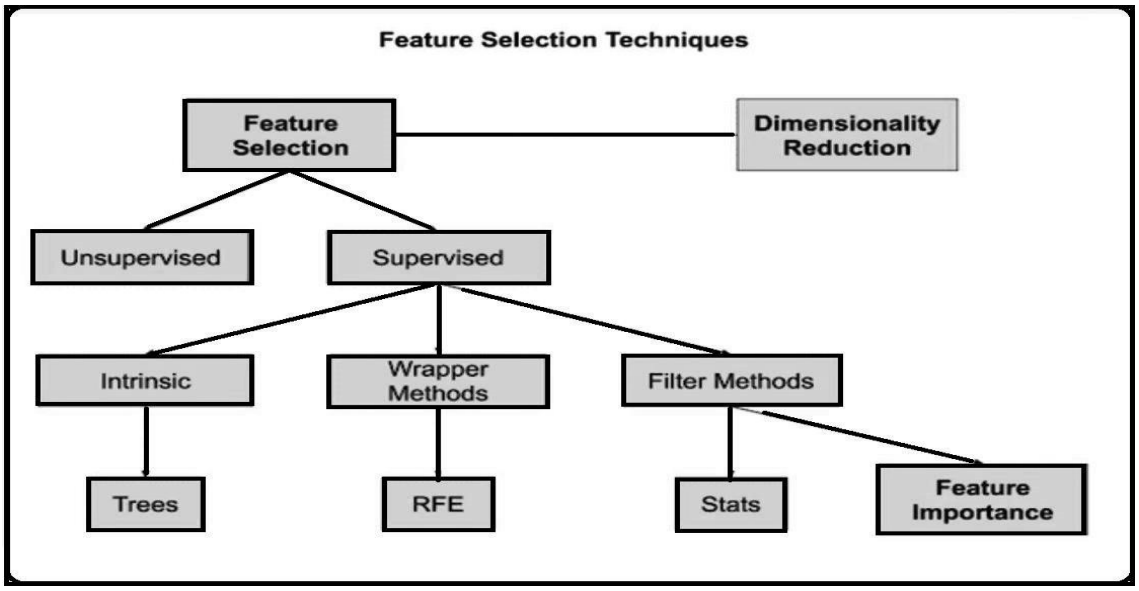

Fig: 1 Feature Selection Technique DFD

4: Engineering

Feature selection refers to the procedure of selecting important features or variables from a dataset's larger available set. By removing unnecessary, redundant, or noisy features from the dataset, feature selection aims to improve a model's accuracy, shorten training time, and prevent overfitting.

By using a feature selection algorithm, data is filtered to remove features that are not pertinent or useful for the target variable. Three steps are typically involved in the feature selection algorithm-based data filtration process:

Feature ranking: In this step, each feature in the dataset is assigned a score or rank based on its relevance to the target variable. Numerous techniques exist for feature ranking, including correlation-based, information-theoretic, and wrapper-based techniques.

Feature selection: In this step, a subset of the top-ranked features is selected for use in the model. The selection of features depends on the ranking method used and the criteria for selecting features, such as a threshold score or a fixed number of features.

Model evaluation: In this stage, the selected subset of features is utilized to evaluate the model's capabilities. The selected features are employed for model training and evaluation if the model performs satisfactorily.If the model performance is not satisfactory, additional feature selection or feature engineering may be necessary.

The main benefit of data filtration by removing unnecessary or redundant features, the feature selection algorithm reduces the dimensionality of the dataset and enhances the performance of the model. Though not always necessary or appropriate for all types of datasets or models, feature selection is an important consideration that should be carefully weighed against the potential advantages. and drawbacks of using feature selection techniques for a given problem.

5: Proposed Model

Model to filter news articles using a feature selection algorithm:

1) Data collection: Collect news articles from various sources.

2) Text preprocessing: Remove all stop words, stem or lemmatize the words, and preprocess the text by making it into a numerical representation, like a Bag of Words or TF-IDF matrix.

Document/ Term	T1	T2	T3	T4	T5	T6
D1	5	9	4	0	5	6
D2	0	8	5	3	10	8
D3	3	5	6	6	5	0
D4	4	6	7	8	4	4

Table 1: Words (T1,T2,T3,T4,T5,T6) Number of times (5,0,3,4) in a Document (D1,D2,D3,D4)

D= Document

T= Particular Word

Numerical Value = Number of Times

$$W_{i,j} = tf_{i,j} \times \log(N/df_i)$$

$tf_{i,j}$ = number of occurrences of i in j

$df_{i,j}$ =number of documents containing i

N = total number of documents

By using the above formula, we can calculate the Term frequency of a word present in a document.

TF — IDF (T1 in D1) = 5 log(4/3) = 0.625

TF – IDF (T2 in D1) = 9 * log (4/4) = 0.0

TF — IDF (T3 in D1) = 4 * log (4/4) = 0.0

TF – IDF (T4 in D1) = 0 * log (4/3) = 0.0

TF – IDF(T5 in D1) = 5 *log(4/4) = 0.0

TF – IDF (T6 in D1) = 6 *log (4/3) = 0.7496

TF — IDF(T1 in D2) = 0 * log (4/3) = 0.0

Let's suppose In this corpus, consisting of two documents, Text A and Text B, we intend to construct a TF-IDF matrix.

- Text A: "Jupiter is the largest planet."

- Text B: "Mars is the fourth planet from the sun."The table below shows the values of TF for A and B, IDF, and TFIDF for A and B.

Words	TF (A)	TF (B)	IDF	TFIDF (A)	TFIDF (B)
jupiter	1/5	0	In (2/1)=0.69	0.138	0
is	1/5	1/8	In (2/2)=0	0	0
the	1/5	2/8	In (2/2)=0	0	0
largest	1/5	0	In (2/1)=0.69	0.138	0
planet	1/5	1/8	In (2/2)=0	0.138	0
mars	1/8	1/8	In (2/1)=0.69	0	0.086
fourth	1/8	1/8	In (2/1)=0.69	0	0.086
from	1/8	1/8	In (2/1)=0.69	0	0.086
sun	1/8	1/8	In (2/1)=0.69	0	0.086

Table 2 (Example)

3) Feature ranking: Rank the features using a feature ranking algorithm, such as mutual information or chi-squared test, to choose the features that matter most for the classification task.

4) Feature selection: Select the top-ranked features based on a predetermined threshold or a fixed number of features.

5) Model training: Train a classifier, using the chosen features and the corresponding class labels, a model, such as a Naive Bayes or Support Vector Machine (SVM) model (e.g., "politics" or "sports").

In order to categorize a point as either negative or positive, it is essential to establish a decision rule. This decision rule can be defined as follows:

$$\rightarrow \rightarrow$$

$$X \cdot w - c \geq 0$$

putting -c as b, we get

$$\rightarrow \rightarrow$$

$$X \cdot w + b \geq 0$$

Hence

$$\rightarrow\rightarrow$$

$$y = \{+1 \text{ if } X.w+b \geq 0$$

$$\rightarrow\rightarrow$$

$$y = \{ -1 \text{ if } X.w + b < 0$$

6) Model evaluation: Utilize a holdout set of news articles that weren't used during training to assess the model's performance. Check the model's precision, recall, accuracy, and F1 score.

7) Model deployment: Deploy the trained model to filter incoming news articles by classifying them into the appropriate category based on the selected features.

8) By filtering news articles using a feature selection algorithm,the model's capability to reduce interference and enhance the accuracy of the classification task is significant. Nevertheless.it's crucial to consistently assess and revise the model to ensure its ongoing effectiveness in filtering news articles from diverse sources and covering various subjects.

6: Conclusion

In conclusion, using a feature selection algorithm to filter news articles can be an effective way to reduce noise and improve the accuracy of the classification task. By ranking and selecting only the most

relevant features, the model can effectively filter out irrelevant or redundant information, making it easier to identify and categorize news articles based on their topic or subject matter.

However, It's worth highlighting that the performance of the feature selection algorithm relies on the quality and appropriateness of the chosen features, in addition to the classifier model's performance. Therefore, it's important to carefully evaluate and fine-tune the model to ensure it is accurate and effective in filtering news articles.

Overall, by using feature selection algorithms in the filtration of news articles, we can create more efficient and accurate models that can be used to categorize and filter news articles based on their topic or subject matter, helping readers to quickly and easily find the news they are interested in.

References

[1] Bradshaw, P. (2017*). The Online Journalism Handbook: Skills to Survive and Thrive in the Digital Age* (2nd ed.). Routledge. https://doi.org/10.4324/9781315761428

[2] Chu, A., Shi, K., & Wong, C. (2014). Prediction of Average and Perceived Polarity in Online Journalism.

[3] Dai, J., Yang, Y., Zeng, Y., Li, Z., Yang, P., & Liu, Y. (2022). The evolutionary game analysis of public opinion on pollution control in the citizen journalism environment. *Water (Switzerland), 14*(23) doi:10.3390/w14233902

[4] Feature Selection Techniques in Machine Learning - Javatpoint. Www.javatpoint.com, https://www.javatpoint.com/feature-selection-techniques-in-machine-learning.

[5] Grove, D, F,.Boghe, K,.&Marez. (2020) (What) Can Journalism Studies Learn from Supervised Machine Learning?,*Journalism Studies*, 21:7, 912-927, DOI: 10.1080/1461670X.2020.1743737

[6] Hintz, A., Treré, E., & Owen, N. (2022). Journalism and data justice: Critically reporting datafication. *The routledge companion to news and journalism* (pp. 179-187) doi:10.4324/9781003174790-22 Retrieved from www.scopus.com

[7] Jacobi, C., Atteveldt, W, V.,& Welbers, K. (2016) Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4:1, 89-106, DOI: 10.1080/21670811.2015.1093271

[8] Li, L. (2022). Data news dissemination strategy for decision making using new media platform. *Soft Computing, 26*(20), 10677-10685. doi:10.1007/s00500-022-06819-0

[9] Lan, Y. -. (2017). A hybrid feature selection based on mutual information and genetic algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 7(1), 214-225. doi:10.11591/ijeecs.v7.i1.pp214-225

[10] Morini, F. (2023). Data journalism as “Terra incognita”: Newcomers’ tensions in shifting towards data journalism epistemology. *Journalism Practice*, doi:10.1080/17512786.2023.2185656

[11] Nanekar, E., Nalawade, S., Castelino, Z., &Rukhande, S. (2023). *Automated journalism based on sports analysis* doi:10.1007/978-3-031-13150-9_45 Retrieved from www.scopus.com

[12] Nag, T., Ghosh, J., Mukherjee, M., Basak, S., & Chakraborty, S. (2023). A python-based virtual AI assistant doi:10.1007/978-981-19-4052-1_58 Retrieved from www.scopus.com

[13] Rahman, A. A., & Bockarie, A. (2022). Exploring trainee characteristics affecting transfer of training: Vocational training of in pakistan. *International Journal of Training Research,* doi: .1080/ 14480220. 2022. 2081241

[14] Singh, T., Kumari, M., Gupta, D. S., & Siniak, N. (2024). Axiomatic analysis of pre-processing methodologies using machine learning in text mining: (pp. 229-256) doi:10.1002/ 9781119905233 .ch11 Retrieved from www.scopus.com

[15] Tran, C. -. (2019). Face recognition based on similarity feature-based selection and classification algorithms and wrapper model. International Journal of Machine Learning and Computing, 9(3), 357-362. doi:10.18178/ijmlc.2019.9.3.810

[16] Tran, D. H., Sheng, Q. Z., Zhang, W. E., Tran, N. H., & Khoa, N. L. D. (2023). CupMar: A deep learning model for personalized news recommendation based on contextual user-profile and multi-aspect article representation. World Wide Web, 26 (2), 713-732. doi: 10.1007/ s11280-022-01059-6

[17] Vijay, S., Mann, P., Chaudhary, R., & Rana, A. (2023). *Emerging trends in multimedia* doi:10.1007/978-981-19-4193-1_29 Retrieved from www.scopus.com

[18] Wu, C., Wu, F., Huang, Y., &Xie, X. (2023). Personalized news recommendation: Methods and challenges. ACM Transactions on Information Systems, 41 (1) doi: .1145/ 3530257

[19] Young, M, L,. Hermida A,.& Fulda J,. (2018) What Makes for Great Data Journalism?, *Journalism Practice*, 12:1, 115-135, DOI: 10.1080/17512786.2016.1270171

[20] Bhoite, S., Ansari, G., Patil, C.H., Thatte, S., Magar, V., Gandhi, K. (2023), "Stock market prediction using Recurrent Neural Network and Long-Short-Term-Memory", ICT Infrastructure and Computing. Lecture Notes in Networks and Systems, vol 520. Springer, Singapore. (Scopus indexing) https://doi.org/10.1007/978-981-19-5331-6_65 Springer LNNS. ISSN: 2367-3370.

[21] Shubham Khndale, Sachin Bhoite (2019), "Campus Placement Analyzer: Using Supervised Machine Learning Algorithms", International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 358-362, 2019, ISSN:-2319–8656, doi: 10.7753/IJCATR0809.1004