BBC news text summarization using machine learning models based on NLP.

1st Nirnoy Chandra Sarker
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
nirnoy.chandra.sarker@g.bracu.ac.bd

2nd Jennifer Abedin
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
jennifer.abedin@g.bracu.ac.bd

3rd Ishrat Jahan
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
ishrat.jahan5@g.bracu.ac.bd

4thAdib Muhammad Amit
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
adib.muhammad.amit@g.bracu.ac.bd

5th Mehnaz Ara Fazal
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
mehnaz.ara.fazal@g.bracu.ac.bd

6thAnnajiat Alim Rasel
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat.alim.rasel@g.bracu.ac.bd

***Abstract*—Over the past ten years, the amount of data on the Internet has grown continuously. As a result, a solution that converts this enormous amount of unprocessed data into meaningful information that the human mind can perceive is required. we embark on a detailed exploration of text summarization through the application of various machine learning models in the period of NLP. This study has categorized news articles into distinct domains including sports, business, politics, tech and entertainment by utilizing Multinomial Naive Bayes, Decision Tree, Random forest,Neural Network, Support Vector Machine, K-Nearest Neighbors and SGD classifiers. Our paper addresses the broader challenge of obtaining valuable information from large amounts of text along with assessing how better these models summarize diverse news contents. Recent improvements in automatic text summarization has inspired us greatly, particularly the use of neural networks and autonomous feature learning. Our goal is to make summarization systems more efficient and effective. This research contributes to the present discourse surrounding information retrieval from textual data as we traverse the integration of machine learning models and NLP for news summarization. These findings shed light on the current landscape and also offer insights into potential avenues for future exploration and improvement within the dynamic intersection of NLP and machine learning for text summarization in the context of BBC news reporting.**

***Index Terms*—Text summarization,Neural Network,TF-IDF,Machine learning,NLP,Stochastic Gradient Descent (SGD) Classifier.**

I. Introduction

In the dynamic environment of Natural Language Processing (NLP), this study delves into an investigation titled "BBC News Text Summarization using Machine Learning Model based on NLP". The complexity of this research unfolds against the backdrop of the information age where an unparalleled volume of data is generated on a regular basis. This data flood is a result of multiple sharing portals, including social networking sites and question-answering platforms. By 2020, it is estimated that approximately 44 zettabytes of data will be generated overall [14].In this research we focused on textual data which is a rich source of information common in all domains like news-wire agencies, research engines, question-answering systems, blogs, e-commerce platforms, digital libraries etc. The valuable information that can be extracted from this textual data is crucial for many applications. For instance, analyzing customer reviews in e-commerce based online shopping helps to improve product features by understanding those analyzed preferences of customers. [9]Also, examining scientific literature reveals insights about cutting-edge research issues, methodologies, solutions, recent outcomes and areas with ongoing obstacles. The primary challenge to deal with this vast amount of textual data is twofold. One is automatic keyword/ Keyphrase extraction (AKE) and another is textual summarization (TS). AKE refers to the instrumental in annotating articles that

reflects their potential categories and relevance to major domains and topics. This process is essential for information recovery that helps to assess the relevance of a data source through extracted keywords. [10]Furthermore, AKE plays an important role in summarization where keywords serve as building blocks for crafting concise summaries. On the contrary, text summarization (TS) focuses on presenting a brief yet all-encompassing overview of a document. It is effective and efficient because TS gives users key information which is a human-readable form. Ultimately, text summarization saves a lot of time. Motivated by the immense possibilities within textual data and the need to simplify information retrieval and consumption, our study examines through the domains of AKE and TS. In this research, both AKE and TS are thoroughly reviewed by a special emphasis on recent advancement . It includes the integration of deep learning approaches. The intersection of these two research problems forms the main focus of our comprehensive study. Our study recognizes itself by employing a diverse set of machine learning models. [11] All of these models contribute a unique perspective to the overarching goal of categorizing news articles from BBC into distinct domains for example business, sports, entertainment, tech and politics. Apart from these categorizations, our paper recognizes the fundamental challenge of extracting valuable information from extensive textual data. At present, advances in making text summarization automatic, particularly the interrogation of neural networks and unsupervised feature learning has served as a guiding light of our exploration. Our goal is not only to evaluate the effectiveness of these machine learning models in summarizing diverse news content but also to contribute to the recent discourse on enhancing the efficiency and quality of summarization systems. [12]As we direct the convergence of machine learning models and NLP for BBC news summarization, our findings aim to gain the current landscape while offering insights into potential opportunities for future research and improvement within this dynamic intersection.

II. Literature Review

The study [1]investigates the crucial role of summarization in managing the overwhelming volume of internet data. Employing a deep-learning algorithm based on the Spacy library, the research specifically delves into news articles, assessing the influence of named entity recognition (NER) on the summarization process. The evaluation, conducted on datasets from CNN-DailyMail and the BBC (entertainment articles), showcases the significant enhancement in recall, precision, and F-score achieved by the proposed NER method compared to conventional word frequency approaches. Particularly noteworthy is its effectiveness in summarizing shorter texts from the BBC. The introduction highlights the challenges posed by information overload in the digital era, underscoring the need for efficient text summarization techniques. In conclusion, the study points to the transformative potential of the NER-based approach, providing a swift and accurate solution for news consumption, thereby contributing to advancements in natural language processing and news summarization.

The study [6] explores three prominent natural language processing models—Transformer, BERT, and PEGASUS—aiming to enhance abstractive text summarization. Detailing the models' architectures, training objectives, and pre-training methods, the research evaluates their performance on datasets like CNN/DailyMail and InShorts using ROUGE and BLEU scores. PEGASUS consistently outperforms, demonstrating superior scores across all metrics. The introduction frames the problem within the context of information overload, emphasizing the need for efficient text summarization. The objectives delineate a comprehensive evaluation of transformer-based models, focusing on BERT and PEGASUS. The background underscores the critical role of abstractive summarization and the evolution of encoder-decoder models. The study's conclusion suggests PEGASUS as a potential leader but acknowledges the trade-off with computational demands. Overall, the research contributes to advancing transformer-based NLP models, offering valuable insights into abstract text summarization's nuances and model-specific strengths and limitation

The paper [5] critically examines the landscape of text summarization, emphasizing the underrepresentation of abstractive methods in prior literature surveys. It positions natural language processing (NLP) as pivotal in addressing the challenges posed by the vast amount of online textual content. The study offers a fourfold contribution, encompassing a comprehensive survey, comparative evaluation, insights into summarization processes, and the provision of datasets and code for reproducibility. Methodologically, the paper adopts stringent selection criteria, focusing on well-cited papers and high-impact journals. The conclusion underscores key findings from the experimental evaluation, emphasizing the impact of fine-tuning on abstractive models and the nuanced performance of extractive approaches. The open issues and future work section identifies crucial areas for exploration, including language-agnostic models and leveraging large language models for semi-supervised learning. Overall, the paper significantly advances the understanding of text summarization, providing a robust evaluation framework and valuable directions for future research.

The paper [3] addresses the critical issue of evaluating machine-generated text summaries, particularly when human-generated reference summaries are unavailable. It introduces a novel metric, SUSWIR, which leverages factors like Semantic Similarity, Redundancy, Relevance, and Bias Avoidance Analysis, based on the original text. The motivation stems from the limitations of traditional metrics like ROUGE and METEOR. The manuscript is well-structured, providing a comprehensive background, clear problem definition, and a detailed description of the proposed metric. Experimental results across five datasets showcase the superiority of SUSWIR in scenarios without human reference summaries. The comparison against ROUGE and METEOR, especially

on the BBC Articles dataset, demonstrates the consistent and reliable performance of SUSWIR. The paper concludes by emphasizing the importance of using informative reference texts for evaluation. Overall, the research makes a significant contribution to text summarization evaluation, offering a valuable alternative metric with practical implications for assessing machine-generated summaries in diverse contexts.

The paper [4] addresses the crucial challenge of low-resource language summarization, proposing a methodology that adapts self-attentive transformer-based models (mBERT, mT5) for this purpose. It introduces urT5, a model tailored for the low-resource language Urdu, and creates a novel baseline dataset. The research objectives focus on mitigating the lack of NLP resources for low-resource languages, aiming to establish competitive baseline resources for Urdu summarization. The contributions include a unique methodology for adapting pre-trained language models, the creation of a substantial summarization dataset for Urdu, and competitive evaluation results. The paper is well-structured, presenting a clear introduction, addressing research objectives, and detailing the methodology and experimental results. The significance of the work lies in its practical implications for low-resource language processing. The future research areas proposed further enrich the scope of the study. Overall, the research makes a valuable contribution to advancing NLP capabilities for low-resource languages.

The paper [2] tackles the challenge of rare words in automatic text summarization, proposing a transformer model modification with a pointer-generator layer and frequency information integration to enhance rare word handling. The hybrid model, combining extractive and abstractive elements, is applied to news summarization. The introduction establishes the importance of summarization and frames the rare word problem effectively. The literature review contextualizes the rarity issue over out-of-vocabulary words. Clear problem definition focuses on rare words and fixed input size limitations. The proposed approach is well-motivated, with results showing improvements in ROUGE scores. The conclusion aptly summarizes findings and suggests future work, providing a valuable contribution to rare-word considerations in text summarization.

III. Methodology

In the figure one we have shown our work flow of this research. Initially we generated our data set and used a variety of preposition techniques.

A. Dataset

As we Previously stated, there is not much research on news text summarization. After research we have chosen to use the data set that is accessible on kaggle. We selected this dataset as it is categorized news articles into distinct domains including sports, business, politics, tech and entertainment. The three main columns are Category,heading and article text preposition is a crucial step in NLP that involves transforming raw text data into a format suitable for analysis or ML models.We have trained data and used to access our model.

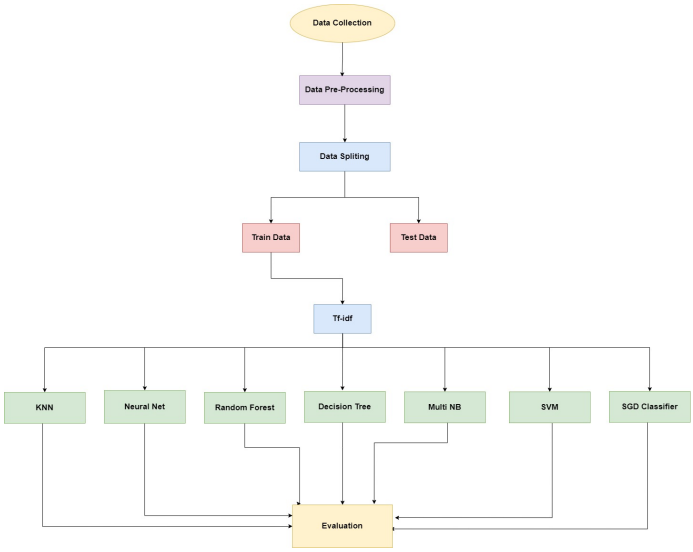

Fig. 1. Multinomial Naive Bayes

TABLE I
DETAILED EVALUATION OF DIFFERENT CLASSIFIER

Category	Count
sport	511
business	510
politics	417
tech	401
entertainment	386

B. Pre processing :

TF-IDF:A text vectorizer called term frequency-inverse document frequency turns text into a useful vector. It combines the notions of document frequency (DF) and term frequency (TF).The quantity of times a certain phrase appears in a text is known as its frequency. Term frequency reveals a term's level of significance inside a document. [7] Term frequency is a matrix that shows all of the data's texts as a row representing the number of documents and a column representing the number of unique words found in all of the documents.According to the theory underpinning TF-IDF, words that we use more frequently in a text are probably more significant to it than terms that we use less frequently. Nonetheless, certain terms, like "the" or "and," are often used in papers but do not provide much information. As a result, the world's inverse document frequency (IDF) is also included in the TF-IDF statistic.

C. Machine Learning Models

A mathematical description of a realworld procedure that is learnt from data is what a machine learning model does. By recognizing patterns and relationships in the data, the model

gains the capacity to project or make justifiable judgments on previously unknown facts.

Multinomial Naive Bayes Classifier: Multinomial Naive Bayes Classifier(Multi NB) is rooted in probability theory which utilizes Bayes theorem to make predictions. It is a naive assumption that feature independence simplifies calculations.It is only Applied in text-related tasks and it mostly handles high-dimensional data by considering word frequencies in documents. Its simplicity and speed make it a popular choice for tasks like spam detection, sentiment analysis, and document categorizations of newspaper articles. [13]It is useful for managing massive amounts of textual data since it is used to classify texts into predetermined classes.In this research not only provides valuable insights into the application of the Multinomial Naive Bayes Classifier in news articles data but also underscores the importance of the reported accuracy of 82.47%. The model's effectiveness in handling categorical data contributes to the broader landscape of machine learning laying the groundwork for potential advancements and applications.

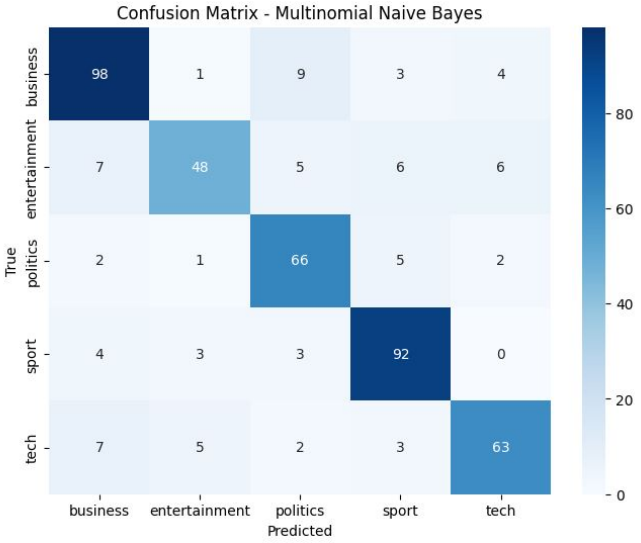

Fig. 2. Multinomial Naive Bayes

Decision Tree Classifier Model: Decision Tree Classifiers create a tree-like structure by recursively splitting datasets based on the most informative features. They are versatile, handling both numerical and categorical data without requiring extensive preprocessing. Their interpretability and intuitive representation of decision-making processes make them valuable for scenarios where transparency and understanding are essential.In order to estimate the grouping of the data set, the decision tree analyzes it. The process starts at the root node of the tree, where it compares the value of the root attribute to the attribute of the record in the real data set. It moves on to the next node, following the branch, based on the comparison. It has 67.42% accuray.The model's effectiveness in handling categorical data contributes to the broader landscape of machine learning models.

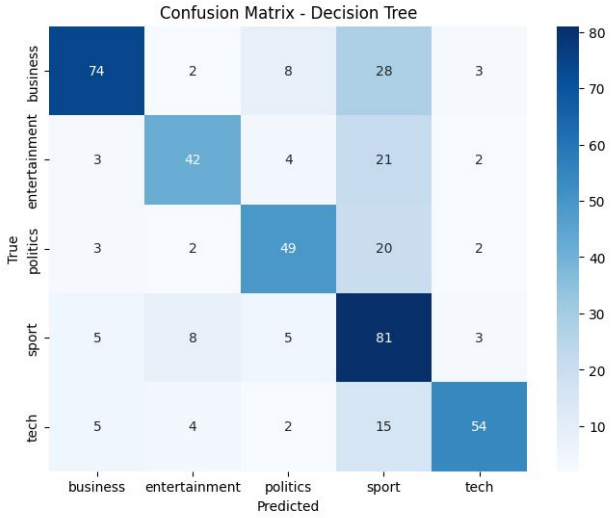

Fig. 3. Decision Tree

Random Forest: Random Forest Classifier extends the decision tree concept by constructing an ensemble of trees randomly. This ensemble approach reduces overfitting. It also provides insights into feature importance, offering robust predictions, especially in scenarios involving high-dimensional data and complex relationships.this research not only provides valuable insights into the application of the Random Forest Classifier in news articles data but also underscores the importance of considering the reported accuracy of 69.66%. The Random Forest's capacity to handle complex relationships in data lays the groundwork for further advancements and applications in the evolving landscape of machine learning and data science.

Support Vector Machine (SVM) : Support Vector Machines (SVM) construct hyperplanes to maximize the margin between different classes in the feature space.SVMs are robust against overfitting and versatile, capable of handling both linear and non-linear scenarios through kernel functions.In this research not only provides valuable insights into the application of the Support Vector Machine model in news articles data but also lays the groundwork for future advancements and applications. The reported accuracy of 81.35% serves as a testament to the efficacy of the SVM in addressing the complexities inherent in the given task, contributing to the ongoing discourse in the realm of machine learning and data science.

Neural Network Classifier: Neural Network Classifiers inspired by the human brain.It consist of interconnected layers of neurons. Deep neural networks with multiple hidden layers

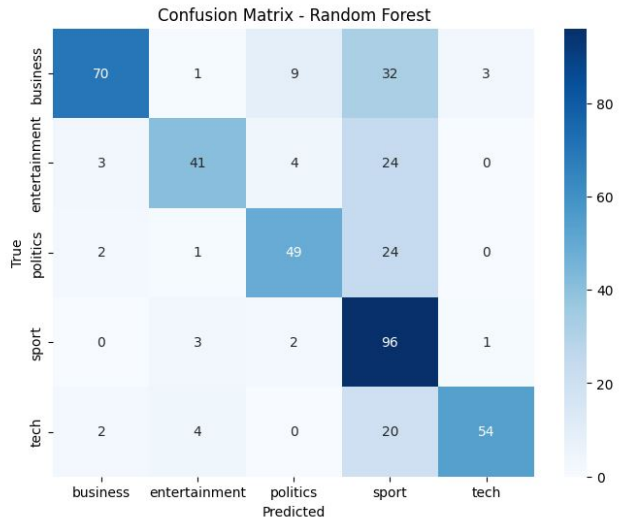

Fig. 4. Random Forest

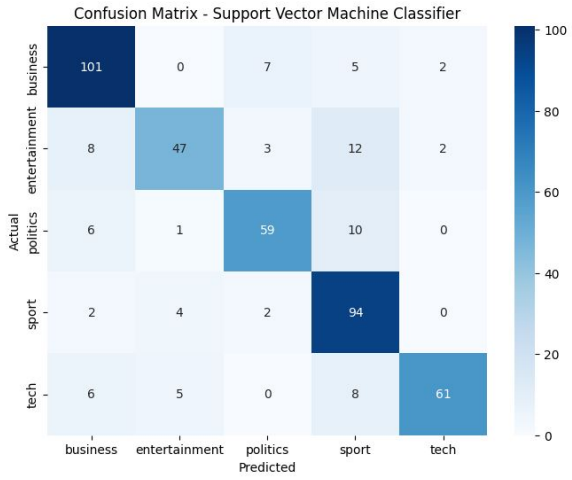

Fig. 5. Support Vector Machine (SVM)

excel at capturing intricate patterns in data. Their adaptability to various data types and automatic feature learning make them powerful tools for tasks like image recognition, natural language processing, and speech recognition.In this research not only provides valuable insights into the application of the Neural Network Classifier in news articles data but also establishes a benchmark with the reported accuracy of 84.27%.Which is highest accuracy above all. The Neural Network's ability to navigate intricate patterns in the data contributes to the ongoing evolution of machine learning and data science, paving the way for future advancements and applications.

Stochastic Gradient Descent (SGD) Classifier: Stochastic Gradient Descent (SGD) Classifiers are optimization algorithms used for training machine learning models.

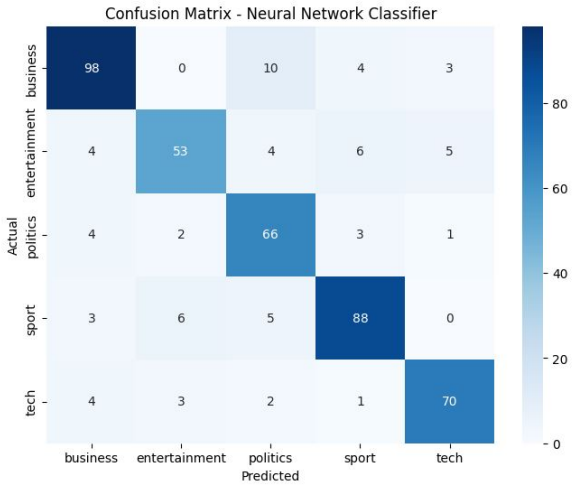

Fig. 6. Neural Network Classifier

Particularly suitable for large-scale tasks, they iteratively update model parameters to minimize errors on the training data. Their efficiency, ease of implementation, and applicability to online learning scenarios where models adapt to new data in real-time contribute to their widespread use in various domains. Each classifier has its unique characteristics, making them valuable tools in different machine learning contexts.In this research not only provides valuable insights into the application of the SGDClassifier in news articles data but also lays the groundwork for future advancements and applications. The reported accuracy of 83.37% serves as a testament to the efficacy of the SGDClassifier in addressing the complexities inherent in the given task, contributing to the ongoing discourse in the realm of machine learning and data science.

.

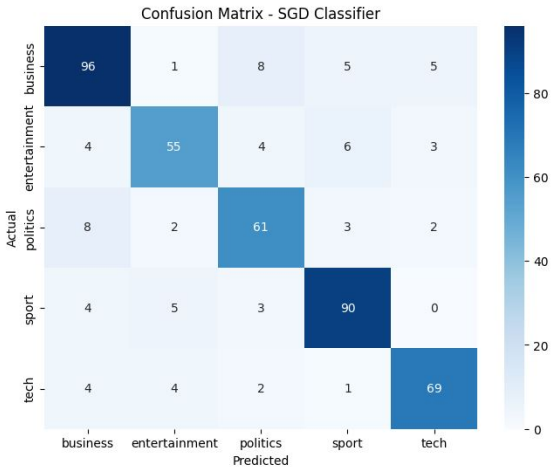

Fig. 7. SGD Classifier

K-Nearest Neighbors Classifier: K-Nearest Neighbors (KNN) Classifiers classify data points based on the majority class among their nearest neighbors. Known for simplicity, they work well with small to moderately sized datasets, making them versatile across different domains, including pattern recognition, image classification, and recommendation systems.In this research not only provides valuable insights into the application of the K-Nearest Neighbors Classifier in news articles data but also lays the groundwork for future advancements and applications. The reported accuracy of 72.81% serves as a testament to the efficacy of the K-Nearest Neighbors in addressing the complexities inherent in the given task, contributing to the ongoing discourse in the realm of machine learning and data science.

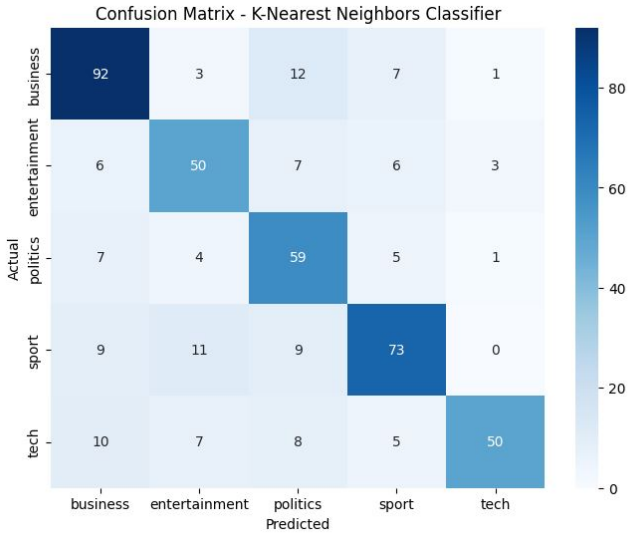

Fig. 8. K-Nearest Neighbors Classifier

IV. EXPERIMENTS AND RESULTS

Numerous thorough studies and experiments have been conducted, including testing and model modification that is regarded as pre-processing. Finding the best pipeline with a variety of improved results—better accuracy and improved performance—was the final goal of this experiment. Through a number of methodical cycles, we systematically identify various pre-processing and model parameters. We employed the TF-IDF approach to remove unnecessary strings, punctuation, and emojis from our Machine Learning (ML) models. Our findings have significantly improved as a consequence of our extensive testing and experimenting. Our pre-processing methods were crucial in raising the caliber of the input data that our machine learning models employed, as seen by the marked shift in the outcomes.The notable shift in the outcomes is sufficient evidence that the pre-processing techniques we employed were crucial in raising the caliber of the input data our machine learning models used. The results of the evaluation we conducted on the models indicate significant differences in their performance. Following extensive testing and experimenting, we discover significant gains in our outcomes. [8]The notable shift in the outcomes is sufficient evidence that the pre-processing techniques we employed were crucial in raising the caliber of the input data our machine learning models used. The results of the evaluation we conducted on the models indicate significant differences in their performance.Our experiment showed that our preprocessing step worked effectively as we found a huge improvement in results in our machine learning models. We have achieved 84.27% accuracy in Neural Network Classifier For a comprehensive breakdown of our model evaluation, please see the detailed table provided below

TABLE II
DETAILED EVALUATION OF DIFFERENT CLASSIFIER

Model Name	Accuracy
Multinomial Naive Bayes Classifier	82.47%
Decision Tree Classifier	67.42%
Random Forest Classifier	69.66%
Neural Network Classifier	84.27%
support Vector Machine	81.35%
K-Nearest Neighbors Classifier	72.8%
SGD Classifier	83.37%

We also can see in a graph:

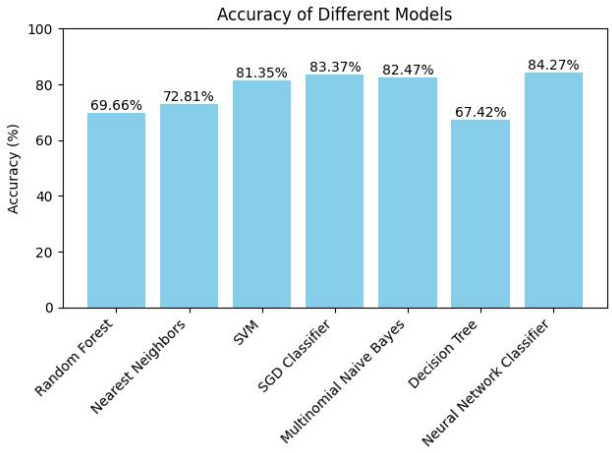

Fig. 9. Models Review

V. CONCLUSION

In conclusion,Our research focuses on the challenges in handling the data from different sources like social networks and news agencies and organizations. We discuss Automatic Keyword Extraction(AKE) for different significant evaluations and categories with a focus on BBC news which is required for effective data processing. AKE creates the stage for Text Summarization, where our machine learning models divide texts into business, sports, entertainment, technology and

governmental categories, applying modern advancements in deep learning. We review the accuracy of these models and suggest possible paths for developments in the developing field of machine learning and natural language processing in news summarization.

REFERENCES

[1] Ibrahim Alshibly, Sabreen Al-Shorfat, Mohammad Otair et al. Text summarization of News Articles based on named entity recognition using Spacy library, 20 March 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2688915/v1]

[2] @articleMOROZOVSKII2023100014, title = Rare words in text summarization, journal = Natural Language Processing Journal, volume = 3, pages = 100014, year = 2023, issn = 2949-7191, doi = https://doi.org/10.1016/j.nlp.2023.100014, url = https://www.sciencedirect.com/science/article/pii/S2949719123000110, author = Danila Morozovskii and Sheela Ramanna

[3] Giarelis, N.; Mastrokostas, C.; Karacapilidis, N. Abstractive vs. Extractive Summarization: An Experimental Review. Appl. Sci. 2023, 13, 7620. https://doi.org/10.3390/ app13137620

[4] Foysal, A.A.; Böck, R. Who Needs External References?—Text Summarization Evaluation Using Original Documents. AI 2023, 4, 970–995. https://doi.org/10.3390/ ai4040049

[5] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1310-1317, doi: 10.1109/ICICT50816.2021.9358703.

[6] C. HARK, T. UÇKAN, E. SEYYARER and A. KARCI, "Graph-Based Suggestion For Text Summarization," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-6, doi: 10.1109/IDAP.2018.8620738.

[7] J. Zenkert, A. Klahold and M. Fathi, "Towards Extractive Text Summarization Using Multidimensional Knowledge Representation," 2018 IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 2018, pp. 0826-0831, doi: 10.1109/EIT.2018.8500186.

[8] T. Islam, M. Hossain and M. F. Arefin, "Comparative Analysis of Different Text Summarization Techniques Using Enhanced Tokenization," 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2021, pp. 1-6, doi: 10.1109/STI53101.2021.9732589.

[9] A. Mishra, A. Sahay, M. a. Pandey and S. S. Routaray, "News text Analysis using Text Summarization and Sentiment Analysis based on NLP," 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, 2023, pp. 28-31, doi: 10.1109/ICSMDI57622.2023.00014.

[10] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

[11] P. Raundale and H. Shekhar, "Analytical study of Text Summarization Techniques," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-4, doi: 10.1109/ASIANCON51346.2021.9544804.

[12] J. Yan and S. Zhou, "A Text Structure-based Extractive And Abstractive Summarization Method," 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2022, pp. 678-681, doi: 10.1109/ICSP54964.2022.9778497.

[13] K. Padmanandam, S. P. V. D. S. Bheri, L. Vegesna and K. Sruthi, "A Speech Recognized Dynamic Word Cloud Visualization for Text Summarization," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 609-613, doi: 10.1109/ICICT50816.2021.9358693.

[14] @articleNASAR2019102088, title = Textual keyword extraction and summarization: State-of-the-art, journal = Information Processing & Management, volume = 56, number = 6, pages = 102088, year = 2019, issn = 0306-4573, doi = https://doi.org/10.1016/j.ipm.2019.102088, url = https://www.sciencedirect.com/science/article/pii/S0306457319300044, author = Zara Nasar and Syed Waqar Jaffry and Muhammad Kamran Malik,